

# An interior penalty discontinuous Galerkin method for a class of monotone quasilinear elliptic problems

Peter W. Fick\*

Faculty of Aerospace Engineering, Delft University of Technology, Delft The Netherlands

January 1, 2014

## Abstract

A family of interior penalty *hp*-discontinuous Galerkin methods is developed and analyzed for the numerical solution of the quasilinear elliptic equation  $-\nabla \cdot (\mathbf{A}(\nabla u) \nabla u) = f$  posed on the open bounded domain  $\Omega \subset \mathbb{R}^d$ ,  $d \geq 2$ . Subject to the assumption that the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}$ ,  $\mathbf{v} \in \mathbb{R}^d$ , is Lipschitz continuous and strongly monotone, it is proved that the proposed method is well-posed. *A priori* error estimates are presented of the error in the broken  $H^1(\Omega)$ -norm, exhibiting precisely the same  $h$ -optimal and mildly  $p$ -suboptimal convergence rates as obtained for the interior penalty approximation of linear elliptic problems. *A priori* estimates for linear functionals of the error and the  $L^2(\Omega)$ -norm of the error are also established and shown to be  $h$ -optimal for a particular member of the proposed family of methods. The analysis is completed under fairly weak conditions on the approximation space, allowing for non-affine and curved elements with multilevel hanging nodes. The theoretical results are verified by numerical experiments.

**Keywords.** *hp*-discontinuous Galerkin methods; interior penalty methods; second-order quasilinear elliptic problems.

## 1 Introduction

Over the past two decades, discontinuous Galerkin (DG) finite element methods have emerged as an effective and popular choice for the numerical solution of a wide range of partial differential equations. This is mainly stimulated by their high degree of locality, their extreme flexibility with respect to *hp*-adaptive mesh refinement, and their natural ability to accommodate high-order discretizations for hyperbolic problems in a locally conservative manner without excessive numerical stabilization. As it stands, there exists a vast amount of literature on the *a priori* error analysis of DG methods for linear problems; we refer to the recent book of Di Pietro & Ern [6] for a comprehensive overview of the most prominent results. For nonlinear problems, however, there are still relatively few results available; we mention the works of Houston *et al.* [15], Ortner & Süli [19], Gudi & Pani [13], Gudi *et al.* [11, 12], [8], Dolejší [7], Bustinza & Gatica [4], and Bi & Lin [3]. It is fair to say that the extension of DG methods from linear to nonlinear problems is non-obvious in many cases, particularly with respect to the proper formulation of the element boundary terms, and that the analysis turns out to be more challenging.

---

\*Email: p.w.fick@tudelft.nl

In this article, we present and analyze a family of interior penalty DG methods for the numerical solution of the following class of quasilinear elliptic boundary value problems. Let  $\Omega$  be an open bounded domain in  $\mathbb{R}^d$ ,  $d \geq 2$ , with Lipschitz boundary  $\partial\Omega = \Gamma_D \cup \Gamma_N$ , where  $\Gamma_D \neq \emptyset$  and  $\Gamma_N = \partial\Omega \setminus \Gamma_D$ . Denoting by  $\mathbf{n}: \Gamma_N \rightarrow \mathbb{R}^d$  the unit outward normal to  $\Gamma_N$ , our model problem of interest is stated as follows: find  $u: \bar{\Omega} \rightarrow \mathbb{R}$  such that

$$\begin{aligned} (1a) \quad & -\nabla \cdot (\mathbf{A}(\mathbf{x}, \nabla u) \nabla u) = f \quad \text{in } \Omega, \\ (1b) \quad & u = g_D \quad \text{on } \Gamma_D, \\ (1c) \quad & \mathbf{A}(\mathbf{x}, \nabla u) \nabla u \cdot \mathbf{n} = g_N \quad \text{on } \Gamma_N, \end{aligned}$$

where  $\mathbf{A} \in [L^\infty(\bar{\Omega} \times \mathbb{R}^d)]^{d,d}$ ,  $f \in L^2(\Omega)$ ,  $g_D \in H^{1/2}(\Gamma_D)$  and  $g_N \in L^2(\Gamma_N)$ . In what follows, we assume that, for  $\mathbf{x} \in \bar{\Omega}$  and  $\mathbf{v} \in \mathbb{R}^d$ , the nonlinear map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{x}, \mathbf{v})\mathbf{v}$  is *Lipschitz continuous* and *strongly monotone*, as phrased by the following statement.

**Assumption 1.1.** *There exist constants  $C_{\mathbf{A}} \geq M_{\mathbf{A}} > 0$  such that, for all  $\mathbf{x} \in \bar{\Omega}$  and all  $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d$ ,*

$$\begin{aligned} (2) \quad & |\mathbf{A}(\mathbf{x}, \mathbf{v}_1)\mathbf{v}_1 - \mathbf{A}(\mathbf{x}, \mathbf{v}_2)\mathbf{v}_2| \leq C_{\mathbf{A}} |\mathbf{v}_1 - \mathbf{v}_2|, \\ (3) \quad & (\mathbf{A}(\mathbf{x}, \mathbf{v}_1)\mathbf{v}_1 - \mathbf{A}(\mathbf{x}, \mathbf{v}_2)\mathbf{v}_2) \cdot (\mathbf{v}_1 - \mathbf{v}_2) \geq M_{\mathbf{A}} |\mathbf{v}_1 - \mathbf{v}_2|^2. \end{aligned}$$

Subject to the above assumption, one can show that problem (1) admits a unique weak solution  $u \in H^1(\Omega)$ . In passing, we note that problems of the type (1) satisfying Assumption 1.1 arise in several applications. A classic example is mean curvature flow, for which  $\mathbf{A}(\mathbf{x}, \nabla u) = (1 + |\nabla u|^2)^{-1/2} \mathbf{I}$  with  $\mathbf{I}$  the  $d \times d$  identity matrix; this has applications in image processing and interface modeling in two-fluid flows, among others. Another example is the modeling of non-Newtonian fluids. For the sake of notational simplicity, we henceforth suppress the dependence of  $\mathbf{A}(\mathbf{x}, \mathbf{v})$  on  $\mathbf{x}$  and simply write  $\mathbf{A}(\mathbf{v})$  instead.

The development of DG methods for problems of the type (1) has also been pursued by several other researchers. In [4], an  $h$ -version local DG method is developed and analyzed exhibiting optimal error estimates in the broken  $H^1(\Omega)$ -norm and  $L^2(\Omega)$ -norm. The development and analysis of  $hp$ -version interior penalty DG methods is initiated by Houston *et al.* [15]. Quasi-optimal error estimates are presented for the error in the broken  $H^1(\Omega)$ -norm, which are optimal in the mesh size  $h$  and mildly supoptimal in the polynomial degree  $p$ , by half an order in  $p$ . Estimates for the error in the  $L^2(\Omega)$ -norm are not presented, but numerical experiments reveal the convergence in the  $L^2(\Omega)$ -norm to be suboptimal. This suboptimality is caused by so-called dual inconsistency of the method due to a particular formulation of the element boundary terms. Difficulties with respect to the proper formulation of the element boundary terms have motivated other researchers to consider the development of *incomplete* interior penalty DG methods; cf. [19, 7, 3]. In [12], a family of interior penalty DG methods is presented and analyzed with a particular choice of the element boundary terms, for which quasi-optimal  $hp$ -error estimates are derived in both the broken  $H^1(\Omega)$ -norm and  $L^2(\Omega)$ -norm.

The purpose of this article is to present and analyze a new family of interior penalty  $hp$ -DG methods for the numerical solution of (1) with quasi-optimal  $hp$ -error estimates in both the broken  $H^1(\Omega)$ -norm and  $L^2(\Omega)$ -norm. As in [15] and [12], our family of methods depends on the parameter  $\theta \in [-1, 1]$ . In the linear setting of  $\mathbf{A}(\cdot) = \mathbf{I}$  with  $\mathbf{I}$  the  $d \times d$  identity matrix and for particular choices of  $\theta$ , the proposed DG formulation reduces to various well-known interior penalty methods; notable examples include the symmetric and nonsymmetric interior penalty methods of, respectively, Arnold [1] and Rivière *et al.* [21]. Subject to Assumption 1.1, we prove that the proposed DG formulation is well-posed provided the discontinuity penalization parameter is chosen sufficiently large. Moreover, a

*a priori* error estimates are presented for the error in the broken  $H^1(\Omega)$ -norm, displaying precisely the same  $h$ -optimal and  $p$ -suboptimal convergence rates as obtained for the interior penalty approximation of linear elliptic problems; cf. [16]. *A priori* estimates for linear functionals of the error and the error in the  $L^2(\Omega)$ -norm are also derived and shown to be  $h$ -optimal when  $\theta = -1$ . The analysis is completed under fairly weak conditions on the  $hp$ -finite element space allowing for non-affine and curved elements with multilevel hanging nodes and non-uniform polynomial degree.

The remainder of this article is organized as follows. Section 2 establishes notation, definitions and some auxiliary results. In Section 3, we introduce the interior penalty  $hp$ -DG approximation of (1) and prove several fundamental properties including a well-posedness result. Section 4 is concerned with the error analysis. Finally, in Section 5 some numerical experiments are presented to illustrate the theoretical results. The appendix is devoted to some auxiliary results regarding the well-posedness of nonlinear variational problems.

## 2 Preliminaries

For  $h > 0$ , let  $\mathcal{T}_h$  be a subdivision of  $\Omega$  into disjoint open element domains  $K$  such that  $\bar{\Omega} = \cup_{K \in \mathcal{T}_h} \bar{K}$ . Here,  $h = \max_{K \in \mathcal{T}_h} h_K$ , where  $h_K = \text{diam}(K)$ . Each  $K \in \mathcal{T}_h$  is the image of a fixed reference domain  $\hat{K}$  under a bijective mapping  $T_K: \hat{K} \rightarrow K$  (that is,  $K = T_K(\hat{K})$  for all  $K \in \mathcal{T}_h$ ), where  $\hat{K}$  is either the open unit simplex or the open unit hypercube in  $\mathbb{R}^d$ . For  $K \in \mathcal{T}_h$ , we denote by  $\mathbf{n}_K$  the unit outward normal with respect to  $\partial K$ . Furthermore, for any pair of neighboring elements  $K, K' \in \mathcal{T}_h$ , we refer to the nonempty  $(d-1)$ -dimensional interior of  $\partial K \cap \partial K'$  as an interior face of  $\mathcal{T}_h$ . Likewise, for any  $K \in \mathcal{T}_h$ , a boundary face lying on  $\Gamma_D$  (resp.  $\Gamma_N$ ) is the nonempty  $(d-1)$ -dimensional interior of  $\partial K \cap \Gamma_D$  (resp.  $\partial K \cap \Gamma_N$ ). The interior faces and the boundary faces lying on  $\Gamma_D$  and  $\Gamma_N$  are collected in the sets  $\mathcal{F}_{h,0}$ ,  $\mathcal{F}_{h,D}$  and  $\mathcal{F}_{h,N}$ , respectively, and we define  $\mathcal{F}_h := \mathcal{F}_{h,0} \cup \mathcal{F}_{h,D} \cup \mathcal{F}_{h,N}$ . In addition, we let  $\mathcal{F}_{h,0,D} := \mathcal{F}_{h,0} \cup \mathcal{F}_{h,D}$ , and, for each  $K \in \mathcal{T}_h$ , we denote by  $\mathcal{F}_{h,K}$  the set of faces lying on  $\partial K$ ; i.e.,  $\mathcal{F}_{h,K} := \{F \in \mathcal{F}_h : F \subset \partial K\}$ . The union of all interior faces is denoted by  $\Gamma_{h,0}$  (i.e.,  $\Gamma_{h,0} := \cup_{F \in \mathcal{F}_{h,0}} F$ ), and analogously we let  $\Gamma_{h,D}$  and  $\Gamma_{h,N}$  represent the union of faces lying on  $\Gamma_D$  and  $\Gamma_N$ . We also define  $\Gamma_{h,0,D} := \Gamma_{h,0} \cup \Gamma_{h,D}$ .

To characterize functions on  $\mathcal{T}_h$  that are possibly discontinuous across inter-element boundaries, we introduce the broken Sobolev space

$$H^s(\Omega, \mathcal{T}_h) := \{v \in L^2(\Omega) : v|_K \in H^s(K), \forall K \in \mathcal{T}_h\},$$

where  $0 < s \leq \infty$ . Here,  $H^s(K)$  denotes the standard Sobolev-Slobodeckij space of order  $s$  for the domain  $K \in \mathcal{T}_h$ . The space  $H^s(\Omega, \mathcal{T}_h)$  is equipped with the broken norm and semi-norm

$$\|v\|_{H^s(\Omega, \mathcal{T}_h)} := \left( \sum_{K \in \mathcal{T}_h} \|v\|_{H^s(K)}^2 \right)^{1/2}, \quad |v|_{H^s(\Omega, \mathcal{T}_h)} := \left( \sum_{K \in \mathcal{T}_h} |v|_{H^s(K)}^2 \right)^{1/2},$$

where  $\|\cdot\|_{H^s(K)}$  and  $|\cdot|_{H^s(K)}$  denote the standard Sobolev-Slobodeckij norm and semi-norm, respectively.

Next, we define jump and average operators for scalar- and vector-valued functions. Let  $K, K' \in \mathcal{T}_h$  be two adjacent element domains sharing an interior face  $F \in \mathcal{F}_{h,0}$ . Given a scalar-valued function  $v \in H^1(\Omega, \mathcal{T}_h)$ , we define the jump and average of  $v$  at  $F$  by

$$[[v]]|_F := v|_K \mathbf{n}_K + v|_{K'} \mathbf{n}_{K'}, \quad \{v\}|_F := (v|_K + v|_{K'})/2.$$

Analogously, for a vector-valued function  $\mathbf{q} \in [H^1(\Omega, \mathcal{T}_h)]^d$ , we set

$$\llbracket \mathbf{q} \rrbracket_F := \mathbf{q}|_K \cdot \mathbf{n}_K + \mathbf{q}|_{K'} \cdot \mathbf{n}_{K'}, \quad \{\!\!\{ \mathbf{q} \}\!\!\}_F := (\mathbf{q}|_K + \mathbf{q}|_{K'})/2.$$

If  $F \in \mathcal{F}_{h,D}$  or  $F \in \mathcal{F}_{h,N}$ , we moreover define  $\llbracket v \rrbracket_F := v|_K \mathbf{n}_K$ ,  $\{\!\!\{ v \}\!\!\}_F := v|_K$  and  $\{\!\!\{ \mathbf{q} \}\!\!\}_F := \mathbf{q}|_K$ , where  $K \in \mathcal{T}_h$  such that  $F \subset \partial K$ ; the quantity  $\llbracket \mathbf{q} \rrbracket_F$  is not required for  $F \in \mathcal{F}_{h,D} \cup \mathcal{F}_{h,N}$  and is thus left undefined.

Given a nonnegative integer  $k$ , let  $\hat{P}_k(\hat{K})$  denote the space of polynomials of total degree up to  $k$  with support on the reference domain  $\hat{K}$ . Also, let  $\hat{Q}_k(\hat{K})$  denote the space of tensor-product polynomials of degree up to  $k$  in each coordinate direction of  $\hat{K}$ . We define  $\hat{S}_k(\hat{K}) = \hat{P}_k(\hat{K})$  when  $\hat{K}$  is the unit  $d$ -simplex, and  $\hat{S}_k(\hat{K}) = \hat{Q}_k(\hat{K})$  when  $\hat{K}$  is the unit  $d$ -hypercube. In addition, let  $S_k(K) = \{v : v \circ T_K \in \hat{S}_k(\hat{K})\}$ . Then, assigning to each  $K \in \mathcal{T}_h$  an integer  $p_K \geq 1$  to represent the local polynomial degree, we introduce the  $hp$ -finite element space

$$V_{h,p} = \{v \in H^1(\Omega, \mathcal{T}_h) : v|_K \in S_{p_K}(K), \forall K \in \mathcal{T}_h\},$$

where  $p = \min_{K \in \mathcal{T}_h} p_K$ .

In the analysis that follows, we make some structural assumptions on the subdivision  $\mathcal{T}_h$  and the distribution of the local polynomial degrees  $\{p_K\}_{K \in \mathcal{T}_h}$ .

**Assumption 2.1.**

- (i) For each  $K \in \mathcal{T}_h$  and some integer  $r_K \geq 2$ , the map  $T_K : \hat{K} \rightarrow K$  is a  $C^{r_K}$ -diffeomorphism satisfying  $|T_K|_{[W_\infty^s(\hat{K})]^{d,d}} \leq \beta_1 h_K^s$  and  $|T_K^{-1}|_{[W_\infty^s(K)]^{d,d}} \leq \beta_1 h_K^{-s}$  for  $s \in [0, r_K]$ , with constant  $\beta_1$  independent of  $h_K$ .
- (ii) The subdivision  $\mathcal{T}_h$  is uniformly graded; i.e., there exists a constant  $\beta_2 > 0$  such that, for all pairs of neighboring elements  $K, K' \in \mathcal{T}_h$  sharing a face  $F \in \mathcal{F}_{h,0}$ , there holds  $\beta_2^{-1} \leq h_K/h_{K'} \leq \beta_2$ .
- (iii) The polynomial degrees  $\{p_K\}_{K \in \mathcal{T}_h}$  have bounded local variation; i.e., there exists a constant  $\beta_3 > 0$  such that, for all pairs of neighboring elements  $K, K' \in \mathcal{T}_h$  sharing a face  $F \in \mathcal{F}_{h,0}$ , there holds  $\beta_3^{-1} \leq p_K/p_{K'} \leq \beta_3$ .

Note that we allow for fairly general subdivisions composed of possibly non-affine and curved elements with multilevel hanging nodes. The only requirement is that each  $K \in \mathcal{T}_h$  is nondegenerate and sufficiently “close” to some affine image of the reference domain  $\hat{K}$  (cf. Assumption 2.1(i); see also, for example, [5]), and that the number of hanging nodes per element face is bounded for all  $K \in \mathcal{T}_h$  (cf. Assumption 2.1(ii)). We remark that, if  $\mathcal{T}_h$  is composed of affine images of simplices and/or multilinear images of hypercubes, then Assumption 2.1(i) reduces to a standard shape regularity condition.

We end this section with some auxiliary results that are needed for the subsequent analysis. Here, and in the sequel, we denote by  $C$  and  $C_i$  ( $i = 1, 2, \dots$ ) generic constants, possibly different on each occurrence, which are independent of  $h$  and  $p$ . In addition, we write  $C \equiv C(\lambda_1, \dots, \lambda_N)$  to indicate the dependence of the constant  $C$  on the parameters  $\lambda_1, \dots, \lambda_N$ . We state without proof the following trace inequality; the proof is analogous to that of Lemma 1.49 in [6].

**Lemma 2.2** (Multiplicative trace inequality). *Let  $K \in \mathcal{T}_h$  and  $F \in \mathcal{F}_{h,K}$ . Then, for any  $v \in H^{s+1}(K)$ ,  $0 \leq s \leq r_K - 1$ , there exists a constant  $C \equiv C(d, \beta_1)$  such that*

$$(4) \quad \|v\|_{H^s(F)}^2 \leq C \left( h_K^{-1} \|v\|_{H^s(K)}^2 + \|v\|_{H^s(K)} \|v\|_{H^{s+1}(K)} \right).$$

For future reference, we also state the following  $hp$ -type inverse estimates; cf. [20, Lemma 3].

**Lemma 2.3** (Inverse estimates). *Let  $K \in \mathcal{T}_h$  and  $F \in \mathcal{F}_{h,K}$ , and denote by  $|K|_d$  and  $|F|_{d-1}$  the corresponding Hausdorff measures of dimension  $d$  and  $d-1$ , respectively. Then, for any  $v \in S_{p_K}(K)$ , there exists a constant  $C \equiv C(d, \beta_1)$  such that:*

(i) for  $0 \leq s \leq r_K - 1$ ,

$$(5) \quad \|v\|_{H^{s+1}(K)} \leq C p_K h_K^{-1/2} \|v\|_{H^s(K)};$$

(ii) for  $0 \leq s \leq r_K$ ,

$$(6) \quad \|v\|_{W_\infty^s(K)} \leq C p_K |K|_d^{-1/2} \|v\|_{H^s(K)},$$

$$(7) \quad \|v\|_{W_\infty^s(F)} \leq C p_K |F|_{d-1}^{-1/2} \|v\|_{H^s(F)}.$$

Using the trace inequality (4) and the inverse estimate (5), and taking into consideration Assumption 2.1, we prove the following result.

**Lemma 2.4.** *Let*

$$(8) \quad \mu_F := \begin{cases} \frac{1}{2}(|K|_d + |K'|_d) / |F|_{d-1} & \text{for } F \in \mathcal{F}_{h,0}, \\ |K|_d / |F|_{d-1} & \text{for } F \in \mathcal{F}_{h,D} \cup \mathcal{F}_{h,N}, \end{cases}$$

and

$$(9) \quad p_F := \begin{cases} \frac{1}{2}(p_K + p_{K'}) & \text{for } F \in \mathcal{F}_{h,0}, \\ p_K & \text{for } F \in \mathcal{F}_{h,D} \cup \mathcal{F}_{h,N}, \end{cases}$$

where  $K, K' \in \mathcal{T}_h$  (resp.  $K \in \mathcal{T}_h$ ) are the element domains adjacent to the face  $F \in \mathcal{F}_{h,0}$  (resp.  $F \in \mathcal{F}_{h,D} \cup \mathcal{F}_{h,N}$ ). There exists a constant  $C \equiv C(d, \beta_1, \beta_2, \beta_3)$  such that, for all  $v \in V_{h,p}$ ,

$$(10) \quad \sum_{F \in \mathcal{F}_h} \frac{\mu_F}{p_F^2} \int_F \{|\nabla v|\}^2 ds \leq C \sum_{K \in \mathcal{T}_h} \int_K |\nabla v|^2 dx.$$

*Proof.* Let  $K \in \mathcal{T}_h$  and  $F \in \mathcal{F}_{h,K}$ . From Assumption 2.1(i) and 2.1(ii) it follows that there exists a constant  $C_1 \equiv C_1(d, \beta_1, \beta_2)$  such that  $\mu_F \leq C_1 h_K$ . Moreover, Assumption 2.1(iii) implies that  $p_F^2 \geq C_2 p_K^2$  for some positive constant  $C_2 \equiv C_2(\beta_3)$ . Hence, by the Young's inequality, we deduce that

$$\sum_{F \in \mathcal{F}_h} \frac{\mu_F}{p_F^2} \int_F \{|\nabla v|\}^2 ds \leq \frac{C_1}{C_2} \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \sum_{F \in \mathcal{F}_{h,K}} \int_F |(\nabla v)|_K|^2 ds.$$

On account of Assumption 2.1(ii) we have that  $\text{card}(\mathcal{F}_{h,K}) \leq C_3$  for some positive integer  $C_3 \equiv C_3(d, \beta_2)$ . Using the trace inequality (4) with constant  $C_4 \equiv C_4(d, \beta_1)$ , we then obtain:

$$\sum_{F \in \mathcal{F}_h} \frac{\mu_F}{p_F^2} \int_F \{|\nabla v|\}^2 ds \leq C_3 C_4 \frac{C_1}{C_2} \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \left( h_K^{-1} \|v\|_{H^1(K)}^2 + \|v\|_{H^1(K)} \|v\|_{H^2(K)} \right).$$

The proof is concluded by applying the inverse estimate (5).  $\square$

### 3 Discontinuous Galerkin method

Let us consider the sum space  $V(h, p) := V_{h,p} + H^s(\Omega)$ ,  $s > 3/2$ . For  $w, v \in V(h, p)$ , we introduce the semilinear form

$$(11) \quad N(w; v) = \sum_{K \in \mathcal{T}_h} \mathbf{A}(\nabla w) \nabla w \cdot \nabla v \, dx + B_0(w; v) + B_D(w; v),$$

and the linear form

$$(12) \quad L(v) = \sum_{K \in \mathcal{T}_h} \int_K f v \, dx + \int_{\Gamma_{h,N}} g_N v \, ds.$$

Here,

$$\begin{aligned} B_0(w; v) = & - \int_{\Gamma_{h,0}} \{ \{ \mathbf{A}(\nabla w - \sigma \llbracket w \rrbracket) \nabla w \} \cdot \llbracket v \rrbracket \, ds \\ & + \theta \int_{\Gamma_{h,0}} \{ \{ \mathbf{A}^T(\nabla w - \sigma \llbracket w \rrbracket) \nabla v \} \cdot \llbracket w \rrbracket \, ds \\ & + \int_{\Gamma_{h,0}} \sigma \{ \{ \mathbf{A}(\nabla w - \sigma \llbracket w \rrbracket) \} \llbracket w \rrbracket \cdot \llbracket v \rrbracket \, ds \\ & + \theta \int_{\Gamma_{h,0}} \sigma^{-1} \{ \{ (\mathbf{A}(\nabla w) - \mathbf{A}(\nabla w - \sigma \llbracket w \rrbracket)) \nabla w \cdot \nabla v \} \, ds \end{aligned}$$

and

$$\begin{aligned} B_D(w; v) = & - \int_{\Gamma_{h,D}} \mathbf{A}(\nabla w - \sigma \mathbf{n}(w - g_D)) \nabla w \cdot \mathbf{n} v \, ds \\ & + \theta \int_{\Gamma_{h,D}} \mathbf{A}^T(\nabla w - \sigma \mathbf{n}(w - g_D)) \nabla v \cdot \mathbf{n}(w - g_D) \, ds \\ & + \int_{\Gamma_{h,D}} \mathbf{A}(\nabla w - \sigma \mathbf{n}(w - g_D)) \mathbf{n} v \cdot \mathbf{n}(w - g_D) \, ds \\ & + \theta \int_{\Gamma_{h,D}} \sigma^{-1} (\mathbf{A}(\nabla w) - \mathbf{A}(\nabla w - \sigma \mathbf{n}(w - g_D))) \nabla w \cdot \nabla v \, ds, \end{aligned}$$

where  $\mathbf{A}^T(\cdot)$  denotes the transpose of  $\mathbf{A}(\cdot)$ ,  $\theta$  is a fixed constant in  $[-1, 1]$ , and  $\sigma$  is a piecewise constant function on  $\Gamma_{h,0,D}$ , defined by

$$\sigma|_F = \alpha \frac{p_F^2}{\mu_F}, \quad F \in \mathcal{F}_{h,0,D}.$$

Here,  $\mu_F$  and  $p_F$  are defined as in (8) and (9), and  $\alpha$  is the so-called *interior penalty parameter*, which is a positive constant independent of  $h$  and  $p$ . As usual, we require that  $\alpha$  is sufficiently large. Anticipating the result of Theorem 3.4, we state that  $\alpha > \alpha_0 = 2C(1 + \lambda_\theta C_{\mathbf{A}}/M_{\mathbf{A}})^2$  will suffice, where  $\lambda_\theta = 1 + |1 + \theta|$  and  $C$  is the constant from Lemma 2.4.

The interior penalty  $hp$ -DG approximation of (1) is now stated as follows: find  $u_{h,p} \in V_{h,p}$  such that

$$(13) \quad N(u_{h,p}; v) = L(v) \quad \forall v \in V_{h,p}.$$

We note that, in the linear case of  $\mathbf{A}(\cdot) = \mathbf{I}$ , with  $\mathbf{I}$  the  $d \times d$  identity matrix, and for particular choices of the parameters  $\theta$  and  $\alpha$ , the DG formulation (13) reduces to various well-known DG methods. Notable examples include the *symmetric* interior penalty method for  $\theta = -1$  and  $\alpha > \alpha_0 > 0$  (cf. [1]), and the *nonsymmetric* interior penalty method for  $\theta = 1$  and  $\alpha > 0$  (cf. [21]).

Under suitable regularity conditions, one can show that (13) is a consistent approximation of (1).

**Lemma 3.1** (Galerkin orthogonality). *Assume that (1) has a strong solution  $u \in H^s(\Omega) \cap C^0(\Omega)$ ,  $s > 3/2$ . Then,*

$$(14) \quad N(u; v) - N(u_{h,p}; v) = 0 \quad \forall v \in V_{h,p}.$$

*Proof.* Since  $u \in C^0(\Omega)$ , we have that  $\llbracket u \rrbracket_F = 0$  strongly for all  $F \in \mathcal{F}_{h,0}$ . Moreover, since  $u$  satisfies (1a) almost everywhere, we have that  $\nabla \cdot \mathbf{A}(\nabla u) \nabla u \in L^2(\Omega)$ . From [6, Lemma 1.24], it then follows that  $\llbracket \mathbf{A}(\nabla u) \nabla u \rrbracket_F = 0$  almost everywhere for all  $F \in \mathcal{F}_{h,0}$ . Therefore, upon integration by parts, we find that  $N(u; v) = L(v)$  for all  $v \in V_{h,p}$ , from which we infer the stated result.  $\square$

For the analysis of the  $hp$ -DG approximation (13), we introduce the norms

$$\begin{aligned} \|v\|^2 &:= \sum_{K \in \mathcal{T}_h} \int_K |\nabla v|^2 dx + \int_{\Gamma_{h,0,D}} \sigma \llbracket v \rrbracket^2 ds, \quad v \in V(h, p), \\ \|v\|_+^2 &:= \|v\|^2 + \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket |\nabla v| \rrbracket^2 ds, \quad v \in V(h, p). \end{aligned}$$

We note that these norms are equivalent on  $V_{h,p}$  for any  $\alpha > 0$ . Indeed, by Lemma 2.4 there exists a constant  $C$  such that

$$(15) \quad \|v\|^2 \leq \|v\|_+^2 \leq (1 + C\alpha^{-1}) \|v\|^2 \quad \forall v \in V_{h,p}.$$

Next, let  $X(\Gamma_{h,0,D}) = \Pi_{K \in \mathcal{T}_h} L^2(\partial K \cap \Gamma_{h,0,D})$  and define the trace operator  $\widehat{\nabla}_\sigma: V(h, p) \rightarrow [X(\Gamma_{h,0,D})]^d$  such that, for  $K \in \mathcal{T}_h$  and  $F \in \mathcal{F}_{h,K}$ ,

$$(\widehat{\nabla}_\sigma w)|_K = \begin{cases} ((\nabla w)|_K)|_F - \sigma \llbracket w \rrbracket & \text{if } F \in \mathcal{F}_{h,0}, \\ ((\nabla w)|_K)|_F - \sigma \mathbf{n}(w|_K - g_D) & \text{if } F \in \mathcal{F}_{h,D}. \end{cases}$$

By the fact that  $\llbracket \llbracket \cdot \rrbracket \rrbracket = \llbracket \cdot \rrbracket$ , we have the following useful identity:

$$\begin{aligned} (16) \quad N(w; v) &= \sum_{K \in \mathcal{T}_h} \mathbf{A}(\nabla w) \nabla w \cdot \nabla v dx \\ &\quad - \int_{\Gamma_{h,0,D}} \llbracket \mathbf{A}(\widehat{\nabla}_\sigma w) \widehat{\nabla}_\sigma w \cdot (\theta \sigma^{-1} \nabla v + \llbracket v \rrbracket) \rrbracket ds \\ &\quad + \theta \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket \mathbf{A}(\nabla w) \nabla w \cdot \nabla v \rrbracket ds. \end{aligned}$$

Rewriting the semilinear form  $N$  according to (16) and using Assumption 1.1, we are able to prove the following two lemmata.

**Lemma 3.2** (Lipschitz continuity). *There exists a constant  $C_N \equiv C_N(\theta, C_{\mathbf{A}})$  such that*

$$(17) \quad N(w_1; v) - N(w_2; v) \leq C_N \|w_1 - w_2\|_+ \|v\|_+ \quad \forall w_1, w_2, v \in V(h, p).$$



*Proof.* Starting from (16) and using that  $|\llbracket \mathbf{q}_1 \cdot \mathbf{q}_2 \rrbracket| \leq \llbracket |\mathbf{q}_1| |\mathbf{q}_2| \rrbracket \leq 2\llbracket |\mathbf{q}_1| \rrbracket \llbracket |\mathbf{q}_2| \rrbracket$  for all  $\mathbf{q}_1, \mathbf{q}_2 \in [H^1(\Omega, \mathcal{T}_h)]^d$ , we have that

$$\begin{aligned} N(w_1; v) - N(w_2; v) &\leq \sum_{K \in \mathcal{T}_h} \int_K |\mathbf{A}(\nabla w_1) \nabla w_1 - \mathbf{A}(\nabla w_2) \nabla w_2| |\nabla v| \, dx \\ &+ \int_{\Gamma_{h,0,D}} \llbracket \mathbf{A}(\widehat{\nabla}_\sigma w_1) \widehat{\nabla}_\sigma w_1 - \mathbf{A}(\widehat{\nabla}_\sigma w_2) \widehat{\nabla}_\sigma w_2 \rrbracket (2|\theta| \sigma^{-1} \llbracket |\nabla v| \rrbracket + \llbracket |v| \rrbracket) \, ds \\ &+ 2|\theta| \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket \mathbf{A}(\nabla w_1) \nabla w_1 - \mathbf{A}(\nabla w_2) \nabla w_2 \rrbracket \llbracket |\nabla v| \rrbracket \, ds. \end{aligned}$$

We use the Lipschitz condition (2) from Assumption 1.1 to bound each of these terms, yielding

$$\begin{aligned} N(w_1; v) - N(w_2; v) &\leq C_{\mathbf{A}} \sum_{K \in \mathcal{T}_h} \int_K |\nabla w_1 - \nabla w_2| |\nabla v| \, dx \\ &+ 2|\theta| C_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \llbracket |w_1 - w_2| \rrbracket \llbracket |\nabla v| \rrbracket \, ds + C_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma \llbracket |w_1 - w_2| \rrbracket \llbracket |v| \rrbracket \, ds \\ &+ 4|\theta| C_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket |\nabla w_1 - \nabla w_2| \rrbracket \llbracket |\nabla v| \rrbracket \, ds. \end{aligned}$$

Upon application of the Cauchy-Schwarz inequality, we arrive at (17) with  $C_N = (2 + 4|\theta|) C_{\mathbf{A}}$ .  $\square$

**Lemma 3.3** (Strong monotonicity). *Let  $\theta \in [-1, 1]$  and select  $\alpha > \alpha_0 = 2C(1 + \lambda_\theta C_{\mathbf{A}}/M_{\mathbf{A}})^2$ , where  $\lambda_\theta = 1 + |1 + \theta|$  and  $C$  is the constant from Lemma 2.4. There exists a positive constant  $M_N \equiv M_N(M_{\mathbf{A}}, \alpha_0/\alpha)$  such that*

$$(18) \quad N(w_1; w_1 - w_2) - N(w_2; w_1 - w_2) \geq M_N \llbracket |w_1 - w_2| \rrbracket^2 \quad \forall w_1, w_2 \in V_{h,p}.$$

*Proof.* Let us write  $w = w_1 - w_2$ . Starting from (16), we have that

$$(19) \quad N(w_1; w_1 - w_2) - N(w_2; w_1 - w_2) = T_1 + T_2 + T_3 + T_4,$$

where

$$\begin{aligned} T_1 &= \sum_{K \in \mathcal{T}_h} \int_K (\mathbf{A}(\nabla w_1) \nabla w_1 - \mathbf{A}(\nabla w_2) \nabla w_2) \cdot \nabla w \, dx, \\ T_2 &= \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket (\mathbf{A}(\widehat{\nabla}_\sigma w_1) \widehat{\nabla}_\sigma w_1 - \mathbf{A}(\widehat{\nabla}_\sigma w_2) \widehat{\nabla}_\sigma w_2) \cdot \widehat{\nabla}_\sigma w \rrbracket \, ds, \\ T_3 &= - (1 + \theta) \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket (\mathbf{A}(\widehat{\nabla}_\sigma w_1) \widehat{\nabla}_\sigma w_1 - \mathbf{A}(\widehat{\nabla}_\sigma w_2) \widehat{\nabla}_\sigma w_2) \cdot \nabla w \rrbracket \, ds, \\ T_4 &= \theta \int_{\Gamma_{h,0,D}} \sigma^{-1} \llbracket (\mathbf{A}(\nabla w_1) \nabla w_1 - \mathbf{A}(\nabla w_2) \nabla w_2) \cdot \nabla w \rrbracket \, ds. \end{aligned}$$

Using the monotonicity condition (3) from Assumption 1.1, it immediately follows that

$$T_1 \geq M_{\mathbf{A}} \sum_{K \in \mathcal{T}_h} \int_K |\nabla w|^2 \, dx.$$



Analogously, for  $T_2$ , we find that

$$\begin{aligned}
T_2 &\geq M_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma^{-1} \{|\widehat{\nabla} w|^2\} \, ds \\
&= M_{\mathbf{A}} \int_{\Gamma_{h,0,D}} (\sigma^{-1} \{|\nabla w|^2\} - 2 \{|\nabla w|\} \cdot \llbracket w \rrbracket + \sigma |\llbracket w \rrbracket|^2) \, ds \\
&\geq -2M_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \{|\nabla w|\} |\llbracket w \rrbracket| \, ds + M_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma |\llbracket w \rrbracket|^2 \, ds.
\end{aligned}$$

The first term on the right hand side can be further bounded by using the Young's inequality  $2ab \leq \epsilon^{-1}a^2 + \epsilon b^2$ , where  $a = \sigma^{-1/2} \{|\nabla w|\}$ ,  $b = \sigma^{1/2} |\llbracket w \rrbracket|$  and  $\epsilon > 0$ . Subsequently applying Lemma 2.4, we obtain

$$\begin{aligned}
T_2 &\geq -M_{\mathbf{A}} \epsilon^{-1} \int_{\Gamma_{h,0,D}} \sigma^{-1} \{|\nabla w|\}^2 \, ds + M_{\mathbf{A}} (1 - \epsilon) \int_{\Gamma_{h,0,D}} \sigma |\llbracket w \rrbracket|^2 \, ds \\
&\geq -M_{\mathbf{A}} \epsilon^{-1} C \alpha^{-1} \sum_{K \in \mathcal{T}_h} \int_K |\nabla w|^2 \, dx + M_{\mathbf{A}} (1 - \epsilon) \int_{\Gamma_{h,0,D}} \sigma |\llbracket w \rrbracket|^2 \, ds,
\end{aligned}$$

where  $C$  is the constant from Lemma 2.4. For  $T_3$ , using the Lipschitz condition (2) from Assumption 1.1 together with the fact that  $\{|\cdot|^2\} \leq 2\{|\cdot|\}^2$ , and proceeding similarly as for  $T_2$ , we have that

$$\begin{aligned}
T_3 &\geq -|1 + \theta| C_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma^{-1} \{|\widehat{\nabla}_{\sigma} w| |\nabla w|\} \, ds \\
&\geq -|1 + \theta| C_{\mathbf{A}} \int_{\Gamma_{h,0,D}} (2\sigma^{-1} \{|\nabla w|\}^2 + |\llbracket w \rrbracket| \{|\nabla w|\}) \, ds \\
&\geq -|1 + \theta| C_{\mathbf{A}} \left( (2 + \epsilon^{-1}) \int_{\Gamma_{h,0,D}} \sigma^{-1} \{|\nabla w|\}^2 \, ds + \epsilon \int_{\Gamma_{h,0,D}} \sigma |\llbracket w \rrbracket|^2 \, ds \right) \\
&\geq -|1 + \theta| C_{\mathbf{A}} \left( (2 + \epsilon^{-1}) C \alpha^{-1} \sum_{K \in \mathcal{T}_h} \int_K |\nabla w|^2 \, dx + \epsilon \int_{\Gamma_{h,0,D}} \sigma |\llbracket w \rrbracket|^2 \, ds \right)
\end{aligned}$$

for any  $\epsilon > 0$ . Finally, for  $T_4$ , using the Lipschitz condition (2) together with the fact that  $|\theta| \leq 1$ , and subsequently applying Lemma 2.4, we obtain

$$\begin{aligned}
T_4 &\geq -|\theta| C_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma^{-1} \{|\nabla w|\}^2 \, ds \\
&\geq -C_{\mathbf{A}} C \alpha^{-1} \sum_{K \in \mathcal{T}_h} \int_K |\nabla w|^2 \, dx.
\end{aligned}$$

Substituting the above bounds for  $T_1$  to  $T_4$  back into (19) and recalling that  $C_{\mathbf{A}} \geq M_{\mathbf{A}} > 0$ , we deduce that

$$\begin{aligned}
&N(w_1; w_1 - w_2) - N(w_2; w_1 - w_2) \\
&\geq (M_{\mathbf{A}} - (2 + \epsilon^{-1}) \lambda_{\theta} C_{\mathbf{A}} C \alpha^{-1}) \sum_{K \in \mathcal{T}_h} \int_K |\nabla w_1 - \nabla w_2|^2 \, dx \\
&\quad + (M_{\mathbf{A}} - \lambda_{\theta} C_{\mathbf{A}} \epsilon) \int_{\Gamma_{h,0,D}} \sigma |\llbracket w_1 - w_2 \rrbracket|^2 \, ds,
\end{aligned}$$

where  $\lambda_\theta = 1 + |1 + \theta|$ . Upon selecting  $\epsilon = M_{\mathbf{A}}/(2\lambda_\theta C_{\mathbf{A}})$ , we arrive at

$$\begin{aligned} N(w_1; w_1 - w_2) - N(w_2; w_1 - w_2) &\geq M_{\mathbf{A}} \left(1 - \frac{\alpha_0}{\alpha}\right) \sum_{K \in \mathcal{T}_h} \int_K |\nabla w_1 - \nabla w_2|^2 dx \\ &\quad + \frac{1}{2} M_{\mathbf{A}} \int_{\Gamma_{h,0,D}} \sigma \|\llbracket w_1 - w_2 \rrbracket\|^2 ds, \end{aligned}$$

where  $\alpha_0 = 2C(1 + \lambda_\theta C_{\mathbf{A}}/M_{\mathbf{A}})^2$ . Hence, we have proved (18) with  $M_N = M_{\mathbf{A}} \max(\frac{1}{2}, 1 - \alpha_0/\alpha)$ . We conclude by noting that  $M_N > 0$  whenever  $\alpha > \alpha_0$ .  $\square$

With the aid of Lemma 3.2 and Lemma 3.3, we are now in the position to prove that the DG approximation (13) admits a unique solution  $u_{h,p} \in V_{h,p}$ . Necessary and sufficient conditions for existence and uniqueness are provided by Theorem A.1 in the Appendix. The following result is an immediate consequence.

**Theorem 3.4** (Existence and uniqueness). *Let  $\theta \in [-1, 1]$  and  $\alpha > \alpha_0 = 2C(1 + \lambda_\theta C_{\mathbf{A}}/M_{\mathbf{A}})^2$ , where  $\lambda_\theta = 1 + |1 + \theta|$  and  $C$  is the constant from Lemma 2.4. Then, the DG approximation (13) has a unique solution  $u_{h,p} \in V_{h,p}$ .*

## 4 A priori error analysis

We begin by introducing the following  $hp$ -approximation results.

**Lemma 4.1.** *Let  $K \in \mathcal{T}_h$  such that  $K = T_K(\hat{K})$ , where  $\hat{K}$  is either the unit  $d$ -simplex or the unit  $d$ -hypercube, and  $T_K$  is a  $C^{r_K}$ -diffeomorphism in compliance with Assumption 2.1(i). For  $s_K \geq 0$ , let  $v \in H^{s_K}(K)$  and define  $t_K = \min(r_K, s_K)$ . Then, for  $p_K = 1, 2, \dots$ , there exists a mapping  $\pi_K: H^{s_K}(K) \rightarrow S_{p_K}(K)$  and a constant  $C$  independent of  $h_K$ ,  $p_K$  and  $v$  such that:*

(i) for  $0 \leq k \leq t_K$ ,

$$\|v - \pi_K(v)\|_{H^k(K)} \leq C \frac{h_K^{\mu_K - k}}{p_K^{t_K - k}} \|v\|_{H^{t_K}(K)};$$

(ii) for  $0 \leq k + 1/2 < t_K$ , and for  $F \in \mathcal{F}_{h,K}$ ,

$$\|v - \pi_K(v)\|_{H^k(F)} \leq C \frac{h_K^{\mu_K - k - 1/2}}{p_K^{t_K - k - 1/2}} \|v\|_{H^{t_K}(K)}.$$

Here,  $\mu_K = \min(p_K + 1, r_K, s_K)$ .

*Proof.* We refer to the proof of Lemma 4.5 in [2] for the case that  $K$  is an affine image of the unit triangle or unit quadrilateral. The generalization to non-affine triangles and quadrilaterals follows *mutatis mutandis* by proceeding similarly as in the proof of Theorem 1 of [5] while making use of [2, Lemma 4.1], and subsequently exploiting Assumption 2.1(i). The argument for simplices and hypercubes of dimension  $d > 2$  is completely analogous.  $\square$

**Corollary 4.2.** *For  $s > 3/2$ , let  $\Pi_{h,p}: H^s(\Omega, \mathcal{T}_h) \rightarrow V_{h,p}$  such that  $\Pi_{h,p}(\cdot)|_K = \pi_K(\cdot)$  for  $K \in \mathcal{T}_h$ , where  $\pi_K$  is the mapping from Lemma 4.1. Moreover, let  $v \in H^s(\Omega, \mathcal{T}_h)$  with  $v|_K \in H^{s_K}(K)$ ,  $s_K \geq s$ ,  $K \in \mathcal{T}_h$ , and select  $\alpha > 0$ . There exists a constant  $C$  such that*

$$\|v - \Pi_{h,p}(v)\|_+ \leq C \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K - 2}}{p_K^{2t_K - 3}} \|v\|_{H^{t_K}(K)}^2 \right)^{1/2},$$

where  $t_K = \min(r_K, s_K)$  and  $\mu_K = \min(p_K + 1, r_K, s_K)$ .

*Proof.* Consider  $K \in \mathcal{T}_h$  and  $F \in \mathcal{F}_{h,K}$ . From Assumption 2.1 it follows that there exists positive constants  $C_1 \equiv C_1(d, \beta_1, \beta_2)$  and  $C_2 \equiv C_2(d, \beta_3)$  such that  $C_1^{-1} h_K \leq \mu_F \leq C_1 h_K$  and  $C_2^{-1} p_K^2 \leq p_F^2 \leq C_2 p_K^2$ . Hence,

$$\alpha C_3^{-1} \frac{p_K^2}{h_K} \leq \sigma|_F \leq \alpha C_3 \frac{p_K^2}{h_K},$$

where  $C_3 = C_2/C_1$ . Accordingly, by Young's inequality, we have that, for  $\eta = v - \Pi_{h,p}(v)$ ,

$$\begin{aligned} |||\eta|||_+^2 &= \sum_{K \in \mathcal{T}_h} \int_K |\nabla \eta|^2 dx + \int_{\Gamma_{h,0,D}} \sigma ||[\![\eta]\!]||^2 ds + \int_{\Gamma_{h,0,D}} \sigma^{-1} \{|\nabla \eta|\}^2 ds \\ &\leq \sum_{K \in \mathcal{T}_h} \left( \|\eta\|_{H^1(K)}^2 + \sum_{F \in \mathcal{F}_{h,K}} \left( 2\alpha C_3 \frac{p_K^2}{h_K} \|\eta\|_{L^2(F)}^2 + \alpha^{-1} C_3 \frac{h_K}{p_K^2} \|\eta\|_{H^1(F)}^2 \right) \right). \end{aligned}$$

Here, in view of the approximation estimates from Lemma 4.1,

$$\begin{aligned} \|\eta\|_{H^1(K)}^2 &\leq C \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-2}} \|v\|_{H^{t_K}(K)}^2, \\ \|\eta\|_{L^2(F)}^2 &\leq C \frac{h_K^{2\mu_K-1}}{p_K^{2t_K-1}} \|v\|_{H^{t_K}(K)}^2, \\ \|\eta\|_{H^1(F)}^2 &\leq C \frac{h_K^{2\mu_K-3}}{p_K^{2t_K-3}} \|v\|_{H^{t_K}(K)}^2. \end{aligned}$$

Hence,

$$|||\eta|||_+^2 \leq C \sum_{K \in \mathcal{T}_h} \left( \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-2}} + \alpha C_3 C_4 \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} + \alpha^{-1} C_3 C_4 \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-1}} \right) \|v\|_{H^{t_K}(K)}^2,$$

where  $C_4 = \max_{K \in \mathcal{T}_h} (\text{card}(\mathcal{F}_{h,K}))$ .  $\square$

Using the  $hp$ -approximation estimate from Corollary 4.2, we prove the following *a priori* error bound.

**Theorem 4.3.** *Let  $u$  denote the solution to (1) and suppose that  $u \in H^s(\Omega) \cap C^0(\Omega)$ ,  $s > 3/2$ , with  $u|_K \in H^{s_K}(K)$ ,  $s_K \geq s$ ,  $K \in \mathcal{T}_h$ . Furthermore, let  $\theta \in [-1, 1]$  and  $\alpha > \alpha_0$ , with  $\alpha_0$  as in Lemma 3.3. Then, denoting by  $u_{h,p} \in V_{h,p}$  the solution to (13), there exists a constant  $C$  such that*

$$(20) \quad |||u - u_{h,p}|||_+ \leq C \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right)^{1/2},$$

where  $t_K = \min(r_K, s_K)$  and  $\mu_K = \min(p_K + 1, r_K, s_K)$ .

*Proof.* Denote by  $\Pi_{h,p}: H^s(\Omega, \mathcal{T}_h) \rightarrow V_{h,p}$  the mapping from Corollary 4.2, and let us write  $u - u_{h,p} = \eta + \xi$ , where  $\eta = u - \Pi_{h,p}(u)$  and  $\xi = \Pi_{h,p}(u) - u_{h,p}$ . Using Lemma 3.3, the Galerkin-orthogonality property (14) and Lemma 3.2, we have that

$$\begin{aligned} M_N |||\xi|||^2 &\leq N(\Pi_{h,p}(u); \xi) - N(u_{h,p}; \xi) \\ &\leq N(\Pi_{h,p}(u); \xi) - N(u; \xi) \\ &\leq C_N |||\eta|||_+ |||\xi|||_+. \end{aligned}$$

Since  $\xi \in V_{h,p}$ , we note from (15) that there exists a constant  $C$  such that  $\|\xi\|_+^2 \leq C \|\xi\|^2$ . Hence,

$$\|\xi\|_+ \leq C \frac{C_N}{M_N} \|\eta\|_+,$$

and therefore, by the triangle inequality,

$$\|u - u_{h,p}\|_+ \leq \|\eta\|_+ + \|\xi\|_+ \leq \left(1 + C \frac{C_N}{M_N}\right) \|\eta\|_+.$$

The estimate (20) then follows by applying Corollary 4.2.  $\square$

We remark that the error estimate obtained in Theorem 4.3 displays the same quasi-optimality as the error estimates obtained for interior penalty DG approximations of linear elliptic problems; cf., for example, [16, Theorem 4.5]. That is, provided that  $r_K \geq s_K \geq p_K + 1$  for all  $K \in \mathcal{T}_h$ , the estimate (20) is optimal in  $h$  and slightly suboptimal in  $p$ , by half an order in  $p$ . Here, the condition that  $r_K \geq s_K$  for all  $K \in \mathcal{T}_h$  reflects the dependence of the estimates on the regularity of the mappings  $\{T_K\}_{K \in \mathcal{T}_h}$ , and stresses the importance of proper mesh design, especially when curved elements are used; cf. [5].

Next, let  $\psi \in L^2(\Omega)$  and consider the linear functional  $J_\psi(w) = (\psi, w)_\Omega$ , where  $w \in V(h, p)$  and  $(\cdot, \cdot)_\Omega$  denotes the  $L^2(\Omega)$  inner product. We shall now be concerned with obtaining a bound for the error  $J_\psi(u) - J_\psi(u_{h,p})$ . The analysis is based on a duality argument and relies on Fréchet differentiability of the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  with respect to  $\mathbf{v}$ . Accordingly, if the limit exists, let us denote by

$$(21) \quad \mathbf{a}'(\mathbf{q}; \mathbf{w}) := \lim_{t \rightarrow 0} \frac{\mathbf{A}(\mathbf{q} + t\mathbf{w})(\mathbf{q} + t\mathbf{w}) - \mathbf{A}(\mathbf{q})\mathbf{q}}{t}, \quad \mathbf{q}, \mathbf{w} \in \mathbb{R}^d,$$

the derivative of the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  at  $\mathbf{q}$  in the direction  $\mathbf{w}$ . Thanks to Assumption 1.1 we are able to make the following claim.

**Lemma 4.4.** *Let  $\mathbf{A}$  satisfy the Lipschitz condition (2) of Assumption 1.1. Then, the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is Fréchet differentiable almost everywhere. That is, for almost every  $\mathbf{q} \in \mathbb{R}^d$ , we have that:*

- (i) *the limit (21) exists for all  $\mathbf{w} \in \mathbb{R}^d$ ;*
- (ii) *the mapping  $\mathbf{w} \mapsto \mathbf{a}'(\mathbf{q}; \mathbf{w}): \mathbb{R}^d \rightarrow \mathbb{R}^d$  is linear and continuous;*
- (iii)  *$\mathbf{a}'(\mathbf{q}; \mathbf{w}) = \mathbf{A}(\mathbf{q} + \mathbf{w})(\mathbf{q} + \mathbf{w}) - \mathbf{A}(\mathbf{q})\mathbf{q} + o(|\mathbf{w}|)$  as  $\mathbf{w} \rightarrow \mathbf{0}$  in  $\mathbb{R}^d$ .*

*Proof.* The lemma is an immediate consequence of Rademacher's Theorem; see, for example, [9, Section 3.1.2].  $\square$

For simplicity of presentation, and without loss of generality, we henceforth assume that the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is *everywhere* Fréchet differentiable in  $\mathbb{R}^d$ , and we refer to Remark 4.8 below for further discussion. Then, for  $\mathbf{q} \in \mathbb{R}^d$ , let  $\mathbf{A}^*(\mathbf{q}) \in \mathbb{R}^{d,d}$  such that  $\mathbf{A}^*(\mathbf{q})\mathbf{v} \cdot \mathbf{w} = \mathbf{a}'(\mathbf{q}; \mathbf{w}) \cdot \mathbf{v}$  for all  $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ . Given  $\psi \in L^2(\Omega)$ , we introduce the dual problem: find  $z: \Omega \rightarrow \mathbb{R}$  such that

$$(22a) \quad -\nabla \cdot (\mathbf{A}^*(\nabla u) \nabla z) = \psi \quad \text{in } \Omega,$$

$$(22b) \quad z = 0 \quad \text{on } \Gamma_D,$$

$$(22c) \quad \mathbf{A}^*(\nabla u) \nabla z \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N.$$

Using Assumption 1.1, it is easy to verify that  $|\mathbf{A}^*(\mathbf{q})\mathbf{v}| \leq C_{\mathbf{A}}|\mathbf{v}|$  and  $\mathbf{A}^*(\mathbf{q})\mathbf{v} \cdot \mathbf{v} \geq M_{\mathbf{A}}|\mathbf{v}|^2$  for all  $\mathbf{q}, \mathbf{v} \in \mathbb{R}^d$ , where  $C_{\mathbf{A}}$  and  $M_{\mathbf{A}}$  are the constants from (2) and (3). Hence, by the Lax-Milgram theorem we deduce that (22) has a unique weak solution  $z \in H^1(\Omega)$ . In what follows, we shall assume slightly stronger regularity by supposing that there exists a strong solution  $z \in H^2(\Omega)$  satisfying

$$(23) \quad \|z\|_{H^2(\Omega)} \leq C\|\psi\|_{L^2(\Omega)}.$$

From [10, Theorem 8.12], we note that this is satisfied if  $\partial\Omega$  is of class  $C^2$  with  $\Gamma_N = \emptyset$ , and if  $\mathbf{A}^*(\nabla u) \in [C^{0,1}(\bar{\Omega})]^{d,d}$ .

With the aid of the dual problem (22) we are able to derive the following *a priori* bound for the error  $J_\psi(u) - J_\psi(u_{h,p})$ .

**Theorem 4.5.** *Consider the same premises as in Theorem 4.3. Furthermore, assume that the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is everywhere Fréchet differentiable in  $\mathbb{R}^d$ , and given  $\psi \in L^2(\Omega)$ , suppose that the dual problem (22) has a strong solution  $z \in H^2(\Omega)$  with  $z|_K \in H^{\ell_K}(K)$ ,  $\ell_K \geq 2$ ,  $K \in \mathcal{T}_h$ . Then, there exists a constant  $C$  such that*

$$(24) \quad \begin{aligned} J_\psi(u) - J_\psi(u_{h,p}) &\leq C \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right)^{1/2} \\ &\quad \times \left( \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\lambda_K-2}}{p_K^{2m_K-3}} \|z\|_{H^{m_K}(K)}^2 \right)^{1/2} + \frac{1+\theta}{\sqrt{\alpha}} \|z\|_{H^2(\Omega)} \right) + R, \end{aligned}$$

where  $t_K = \min(r_K, s_K)$ ,  $m_K = \min(r_K, \ell_K)$ ,  $\mu_K = \min(p_K + 1, r_K, s_K)$ ,  $\lambda_K = \min(p_K + 1, r_K, \ell_K)$ , and where  $R = o(\|u - u_{h,p}\|) \|z\|_{H^2(\Omega)}$ . Moreover, if the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is twice continuously differentiable everywhere in  $\mathbb{R}^d$ , then there exists a constant  $C$  such that

$$(25) \quad R \leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K^{3/2}}{h_K^{d/2}} \right) \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right) \|z\|_{H^2(\Omega)}.$$

Before we embark on the proof of Theorem 4.5, we first introduce an auxiliary result. By our assumption that the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is everywhere Fréchet differentiable in  $\mathbb{R}^d$ , we have that the map  $y \mapsto N(y; v): V(h, p) \rightarrow \mathbb{R}$  is everywhere Fréchet differentiable in  $V(h, p)$ . Accordingly, for any  $v \in V(h, p)$ , let  $N'(q; w, v)$  denote the derivative of the map  $y \mapsto N(y; v): V(h, p) \rightarrow \mathbb{R}$  at  $q$  in the direction  $w$ , given by

$$N'(q; w, v) = \lim_{t \rightarrow 0} \frac{N(q + tw; v) - N(q; v)}{t}, \quad q, w, v \in V(h, p).$$

We introduce the following auxiliary result.

**Lemma 4.6.** *Let  $u \in H^s(\Omega) \cap C^0(\Omega)$ ,  $s > 3/2$ , denote the solution of (1), and suppose that the dual problem (22) has a strong solution  $z \in H^2(\Omega)$ . Then,*

$$J_\psi(w) = N'(u; w, z) - (1 + \theta) \int_{\Gamma_{h,0,D}} \mathbf{a}'(\nabla u; \llbracket w \rrbracket) \cdot \nabla z \, ds \quad \forall w \in V(h, p).$$

*Proof.* Since  $z \in H^2(\Omega)$ , we have that  $\llbracket z \rrbracket|_F = \mathbf{0}$  for all  $F \in \mathcal{F}_{h,0}$ . Accordingly, evaluating  $N'(u; w, z)$  for any  $w \in V(h, p)$ , we find that

$$(26) \quad N'(u; w, z) = \sum_{K \in \mathcal{T}_h} \int_K \mathbf{a}'(\nabla u; \nabla w) \cdot \nabla z \, dx + \theta \int_{\Gamma_{h,0,D}} \{ \mathbf{a}'(\nabla u; \llbracket w \rrbracket) \cdot \nabla z \} \, ds.$$

Using the dual problem (22) and applying integration-by-parts, we also find that, for all  $w \in V(h, p)$ ,

$$(27) \quad \begin{aligned} J_\psi(w) &= - \sum_{K \in \mathcal{T}_h} \int_K w \nabla \cdot (\mathbf{A}^*(\nabla u) \nabla z) \, dx \\ &= \sum_{K \in \mathcal{T}_h} \left( \int_K \mathbf{A}^*(\nabla u) \nabla z \cdot \nabla w \, dx - \int_{\partial K} \mathbf{A}^*(\nabla u) \nabla z \cdot \mathbf{n}_K w \, ds \right) \\ &= \sum_{K \in \mathcal{T}_h} \int_K \mathbf{A}^*(\nabla u) \nabla z \cdot \nabla w \, dx - \int_{\Gamma_{h,0}} \{ \mathbf{A}^*(\nabla u) \nabla z \} \{ w \} \, ds \\ &\quad - \int_{\Gamma_{h,0,D}} \{ \mathbf{A}^*(\nabla u) \nabla z \} \cdot \llbracket w \rrbracket \, ds. \end{aligned}$$

By [6, Lemma 1.24], it follows that  $\llbracket \mathbf{A}^*(\nabla u) \nabla z \rrbracket|_F = 0$  weakly for all  $F \in \mathcal{F}_{h,0}$ . Thence, comparing (26) and (27) while noting that  $\mathbf{A}^*(\nabla u) \nabla z \cdot \mathbf{w} = \mathbf{a}'(\nabla u; \mathbf{w}) \cdot \nabla z$  for all  $\mathbf{w} \in \mathbb{R}^d$ , we obtain the stated result.  $\square$

With the aid of Lemma 4.6, we now present a proof of Theorem 4.5.

*Proof of Theorem 4.5.* Denote by  $\Pi_{h,p}: H^s(\Omega, \mathcal{T}_h) \rightarrow V_{h,p}$ ,  $s > 3/2$ , the mapping from Corollary 4.2, and let us write  $e = u - u_{h,p}$ . Lemma 4.6 implies that

$$(28) \quad \begin{aligned} J_\psi(u) - J_\psi(u_{h,p}) &= N'(u; e, z) - (1 + \theta) \int_{\Gamma_{h,0,D}} \mathbf{a}'(\nabla u; \llbracket e \rrbracket) \cdot \nabla z \, ds \\ &= N'(u; e, z - \Pi_{h,p}(z)) - (1 + \theta) \int_{\Gamma_{h,0,D}} \mathbf{a}'(\nabla u; \llbracket e \rrbracket) \cdot \nabla z \, ds \\ &\quad + N'(u; e, \Pi_{h,p}(z)). \end{aligned}$$

Considering the first term in (28), we deduce by Lemma 3.2 that

$$\begin{aligned} N'(u; e, z - \Pi_{h,p}(z)) &= \lim_{t \rightarrow 0} \frac{N(u + te; z - \Pi_{h,p}(z)) - N(u; z - \Pi_{h,p}(z))}{t} \\ &\leq \sup_{t > 0} \frac{N(u + te; z - \Pi_{h,p}(z)) - N(u; z - \Pi_{h,p}(z))}{t} \\ &\leq C_N \|\llbracket e \rrbracket\|_+ \|\llbracket z - \Pi_{h,p}(z) \rrbracket\|_+, \end{aligned}$$

where  $C_N$  is the constant from Lemma 3.2. Using the error estimate of Theorem 4.3 and the approximation estimate of Corollary 4.2, we then obtain:

$$\begin{aligned} &N'(u; e, z - \Pi_{h,p}(z)) \\ &\leq C \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right)^{1/2} \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\lambda_K-2}}{p_K^{2m_K-3}} \|z\|_{H^{m_K}(K)}^2 \right)^{1/2}. \end{aligned}$$

Next, applying the Cauchy-Schwarz inequality to the second term in (28), we have that

$$\begin{aligned} (1 + \theta) \int_{\Gamma_{h,0,D}} \mathbf{a}'(\nabla u; \llbracket u - u_{h,p} \rrbracket) \cdot \nabla z \, ds \\ \leq (1 + \theta) \left( \int_{\Gamma_{h,0,D}} \sigma |\mathbf{a}'(\nabla u; \llbracket u - u_{h,p} \rrbracket)|^2 \, ds \right)^{1/2} \left( \int_{\Gamma_{h,0,D}} \sigma^{-1} |\nabla z|^2 \, ds \right)^{1/2}. \end{aligned}$$

Using that  $|\mathbf{a}'(\mathbf{q}; \mathbf{w})| \leq C_{\mathbf{A}} |\mathbf{w}|$  for all  $\mathbf{q}, \mathbf{w} \in \mathbb{R}^d$  and subsequently applying Theorem 4.3, we find:

$$\int_{\Gamma_{h,0,D}} \sigma |\mathbf{a}'(\nabla u; \llbracket e \rrbracket)|^2 \, ds \leq C_{\mathbf{A}} \|\llbracket e \rrbracket\|^2 \leq C \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right).$$

Moreover, arguing similarly as in the proof of Lemma 2.4 and subsequently applying the trace inequality from Lemma 2.2, we deduce that

$$\begin{aligned} \int_{\Gamma_{h,0,D}} \sigma^{-1} |\nabla z|^2 \, ds &\leq C \alpha^{-1} \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \int_{\partial K} |\nabla z|^2 \, ds \\ &\leq C \alpha^{-1} \sum_{K \in \mathcal{T}_h} \frac{h_K}{p_K^2} \left( h_K^{-1} \|z\|_{H^1(K)}^2 + \|z\|_{H^1(K)} \|z\|_{H^2(K)} \right) \\ (29) \quad &\leq C \alpha^{-1} \|z\|_{H^2(\Omega)}^2. \end{aligned}$$

Hence, we obtain:

$$(1 + \theta) \int_{\Gamma_{h,0,D}} \mathbf{a}'(\nabla u; \llbracket e \rrbracket) \cdot \nabla z \, ds \leq C \frac{1 + \theta}{\sqrt{\alpha}} \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right)^{1/2} \|z\|_{H^2(\Omega)}.$$

Substituting the above bounds back into (28), we arrive at the stated estimate (24) with  $R = N'(u; e, \Pi_{h,p}(z))$ .

We claim that  $R = o(\|\llbracket e \rrbracket_+\| \|z\|_{H^2(\Omega)})$ . Fréchet differentiability of the map  $y \mapsto N(y; v): V(h, p) \rightarrow \mathbb{R}$  everywhere in  $V(h, p)$  implies that

$$N'(q; w, v) = N(q + w; v) - N(q; v) + o(\|\llbracket w \rrbracket_+\| \|v\|_+) \quad \text{as } \|\llbracket w \rrbracket_+ \rightarrow 0,$$

for all  $q, w, v \in V(h, p)$ . Hence, by the Galerkin-orthogonality property of Lemma 3.1, we obtain that

$$\begin{aligned} R = N'(u; e, \Pi_{h,p}(z)) &= N(u; \Pi_{h,p}(z)) - N(u_{h,p}; \Pi_{h,p}(z)) + o(\|\llbracket e \rrbracket_+\| \|\Pi_{h,p}(z)\|_+) \\ &= o(\|\llbracket e \rrbracket_+\| \|\Pi_{h,p}(z)\|_+) \end{aligned}$$

as  $\|\llbracket e \rrbracket_+ \rightarrow 0$ . Here, in view of (29), we have that  $\|z\|_+ \leq C \|z\|_{H^2(\Omega)}$ , so that, by the triangle inequality and Corollary 4.2,

$$(30) \quad \|\Pi_{h,p}(z)\|_+ \leq C \|z\|_{H^2(\Omega)}.$$

Therefore, we find that  $R = o(\|\llbracket e \rrbracket_+\| \|z\|_{H^2(\Omega)})$ , as claimed.

It remains to prove the estimate (25) subject to the condition that the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  is twice continuously differentiable everywhere in  $\mathbb{R}^d$ . Accordingly, let

$$\mathbf{a}''(\mathbf{q}; \mathbf{w}_1, \mathbf{w}_2) := \lim_{t \rightarrow 0} \frac{\mathbf{a}'(\mathbf{q} + t\mathbf{w}_2; \mathbf{w}_1) - \mathbf{a}'(\mathbf{q}; \mathbf{w}_1)}{t}, \quad \mathbf{q}, \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d,$$



denote the second-order derivative of the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v})\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  at  $\mathbf{q} \in \mathbb{R}^d$  in the direction  $(\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^d \times \mathbb{R}^d$ , and let there be a constant  $C'_\mathbf{A}$  such that  $|\mathbf{a}''(\mathbf{q}; \mathbf{w}_1, \mathbf{w}_2)| \leq C'_\mathbf{A} |\mathbf{w}_1| |\mathbf{w}_2|$  for all  $\mathbf{q}, \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ . By Taylor's Theorem, we have that

$$(31) \quad \mathbf{A}(\mathbf{v}_1)\mathbf{v}_1 - \mathbf{A}(\mathbf{v}_2)\mathbf{v}_2 = \mathbf{a}'(\mathbf{v}_2; \mathbf{v}_1 - \mathbf{v}_2) + \mathbf{r}(\mathbf{v}_2, \mathbf{v}_1 - \mathbf{v}_2), \quad \forall \mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^d,$$

with the integral remainder

$$\mathbf{r}(\mathbf{v}_2, \mathbf{v}_1 - \mathbf{v}_2) = \int_0^1 \mathbf{a}''(\mathbf{v}_2 + t(\mathbf{v}_1 - \mathbf{v}_2); \mathbf{v}_1 - \mathbf{v}_2, \mathbf{v}_1 - \mathbf{v}_2)(1-t) dt,$$

satisfying  $|\mathbf{r}(\mathbf{v}_2, \mathbf{v}_1 - \mathbf{v}_2)| \leq C'_\mathbf{A} |\mathbf{v}_1 - \mathbf{v}_2|^2$ . Now, recall that  $R = N'(u; e, \Pi_{h,p}(z))$ . Using the Galerkin-orthogonality property of Lemma 3.1 and the Taylor expansion (31), we deduce that

$$\begin{aligned} R &= N'(u; e, \Pi_{h,p}(z)) - N(u; \Pi_{h,p}(z)) + N(u_{h,p}; \Pi_{h,p}(z)) \\ &= - \sum_{K \in \mathcal{T}_h} \int_K \mathbf{r}(\nabla u; \nabla e) \cdot \nabla(\Pi_{h,p}(z)) dx \\ &\quad + \int_{\Gamma_{h,0,D}} \{ \mathbf{r}(\nabla u, \widehat{\nabla}_\sigma e) \cdot (\theta \sigma^{-1} \nabla(\Pi_{h,p}(z)) + \llbracket \Pi_{h,p}(z) \rrbracket) \} ds \\ &\quad - \theta \int_{\Gamma_{h,0,D}} \sigma^{-1} \{ \mathbf{r}(\nabla u, \nabla e) \cdot \nabla(\Pi_{h,p}(z)) \} ds \\ &\leq C'_\mathbf{A} \sum_K \int_K |\nabla e|^2 |\nabla(\Pi_{h,p}(z))| dx \\ &\quad + C'_\mathbf{A} \int_{\Gamma_{h,0,D}} \{ |\widehat{\nabla}_\sigma e|^2 (|\theta| \sigma^{-1} |\nabla(\Pi_{h,p}(z))| + \llbracket \Pi_{h,p}(z) \rrbracket) \} ds \\ &\quad + C'_\mathbf{A} |\theta| \int_{\Gamma_{h,0,D}} \sigma^{-1} \{ |\nabla e|^2 |\nabla(\Pi_{h,p}(z))| \} ds. \end{aligned}$$

By Young's inequality and the fact that  $|\{ \mathbf{q}_1 \cdot \mathbf{q}_2 \}| \leq \{ |\mathbf{q}_1| |\mathbf{q}_2| \} \leq 2 \{ |\mathbf{q}_1| \} \{ |\mathbf{q}_2| \}$  for all  $\mathbf{q}_1, \mathbf{q}_2 \in [H^1(\Omega, \mathcal{T}_h)]^d$ , we then obtain:

$$\begin{aligned} R &\leq C'_\mathbf{A} \sum_K \int_K |\nabla e|^2 |\nabla(\Pi_{h,p}(z))| dx \\ &\quad + 12|\theta| C'_\mathbf{A} \int_{\Gamma_{h,0,D}} (\sigma^{-1} \{ |\nabla e| \}^2 + \sigma \{ |e| \}^2) \{ |\nabla(\Pi_{h,p}(z))| \} ds \\ &\quad + 4 C'_\mathbf{A} \int_{\Gamma_{h,0,D}} (\{ |\nabla e| \}^2 + \sigma^2 \{ |e| \}^2) \llbracket \Pi_{h,p}(z) \rrbracket ds \\ (32) \quad &\leq (4 + 12|\theta|) C'_\mathbf{A} \|e\|_+^2 \|\Pi_{h,p}(z)\|_\star, \end{aligned}$$

where

$$\begin{aligned} \|\Pi_{h,p}(z)\|_\star &= \max_{K \in \mathcal{T}_h} \|\Pi_{h,p}(z)\|_{W_\infty^1(K)} + \max_{F \in \mathcal{F}_{h,0,D}} \{ \{ |\nabla(\Pi_{h,p}(z))| \} \}_{L^\infty(F)} \\ (33) \quad &\quad + \max_{F \in \mathcal{F}_{h,0,D}} \sigma \{ \llbracket \Pi_{h,p}(z) \rrbracket \}_{L^\infty(F)}. \end{aligned}$$

An upper bound for  $\|e\|_+$  is provided by Theorem 4.3. To prove (25), it thus remains to show that  $\|\Pi_{h,p}(z)\|_\star \leq C \max_{K \in \mathcal{T}_h} \left( p_K^{3/2} h_K^{-d/2} \right) \|z\|_{H^2(\Omega)}$ . To this end, let us note that,

in view of Lemma 4.1 and the triangle inequality, there exists a constant  $C$  such that  $\|\Pi_{h,p}(z)\|_{H^2(K)} \leq C\|z\|_{H^2(K)}$ . Thence, exploiting the inverse estimate (6), we have that

$$\begin{aligned} \max_{K \in \mathcal{T}_h} \|\Pi_{h,p}(z)\|_{W_\infty^1(K)} &\leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K}{h_K^{d/2}} \|\Pi_{h,p}(z)\|_{H^1(K)} \right) \\ &\leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K}{h_K^{d/2}} \right) \|z\|_{H^2(\Omega)}. \end{aligned}$$

For the second term in (33), we apply the inverse estimate (7) to obtain

$$\begin{aligned} \max_{F \in \mathcal{F}_{h,0,D}} \left\| \llbracket |\nabla(\Pi_{h,p}(z))| \rrbracket \right\|_{L^\infty(F)} &\leq \max_{K \in \mathcal{T}_h} \left( \max_{F \in \mathcal{F}_{h,K}} \left\| (\Pi_{h,p}(z))|_K \right\|_{W_\infty^1(F)} \right) \\ &\leq C \max_{K \in \mathcal{T}_h} \left( \max_{F \in \mathcal{F}_{h,K}} \frac{p_K}{(|F|_{d-1})^{1/2}} \left\| (\Pi_{h,p}(z))|_K \right\|_{H^1(F)} \right). \end{aligned}$$

On account of Assumption 2.1, there exists a constant  $C \equiv C(d, \beta_1, \beta_2)$  such that  $|F|_{d-1} \geq C h_K^{d-1}$  for all  $F \in \mathcal{F}_{h,K}$ ,  $K \in \mathcal{T}_h$ . Applying the trace inequality (4), we then find that

$$\begin{aligned} \max_{F \in \mathcal{F}_{h,0,D}} \left\| \llbracket |\nabla(\Pi_{h,p}(z))| \rrbracket \right\|_{L^\infty(F)} &\leq C \max_{K \in \mathcal{T}_h} \frac{p_K}{h_K^{d/2}} \left( \|\Pi_{h,p}(z)\|_{H^1(K)}^2 + h_K \|\Pi_{h,p}(z)\|_{H^1(K)} \|\Pi_{h,p}(z)\|_{H^2(K)} \right)^{1/2} \\ &\leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K}{h_K^{d/2}} \right) \|z\|_{H^2(\Omega)}. \end{aligned}$$

Finally, considering the third term in (33), we deduce that, by Assumption 2.1 and the inverse estimate (7),

$$\begin{aligned} \max_{F \in \mathcal{F}_{h,0,D}} \sigma \left\| \llbracket \Pi_{h,p}(z) \rrbracket \right\|_{L^\infty(F)} &= \max_{K \in \mathcal{T}_h} \left( \max_{F \in \mathcal{F}_{h,K} \cap \mathcal{F}_{h,0,D}} \sigma \left\| \llbracket \Pi_{h,p}(z) \rrbracket \right\|_{L^\infty(F)} \right) \\ &\leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K^3}{h_K^{(d+1)/2}} \max_{F \in \mathcal{F}_{h,K} \cap \mathcal{F}_{h,0,D}} \left\| \llbracket \Pi_{h,p}(z) \rrbracket \right\|_{L^2(F)} \right). \end{aligned}$$

By the fact that  $z \in H^1(\Omega)$  with  $z = 0$  on  $\Gamma_D$ , we have that  $\left\| \llbracket \Pi_{h,p}(z) \rrbracket \right\|_{L^2(F)} = \left\| \llbracket z - \Pi_{h,p}(z) \rrbracket \right\|_{L^2(F)}$  for all  $F \in \mathcal{F}_{h,0,D}$ . Applying Lemma 4.1, we then obtain:

$$\begin{aligned} \max_{F \in \mathcal{F}_{h,0,D}} \sigma \left\| \llbracket \Pi_{h,p}(z) \rrbracket \right\|_{L^\infty(F)} &\leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K^3}{h_K^{(d+1)/2}} \max_{F \in \mathcal{F}_{h,K}} \|z - (\Pi_{h,p}(z))|_K\|_{L^2(F)} \right) \\ &\leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K^{3/2}}{h_K^{(d-2)/2}} \right) \|z\|_{H^2(\Omega)}. \end{aligned}$$

Substituting the above inequalities back into (33), we thus find that  $\|\llbracket \Pi_{h,p}(z) \rrbracket\|_* \leq C \max_{K \in \mathcal{T}_h} (p_K^{3/2} h_K^{-d/2}) \|z\|_{H^2(\Omega)}$  which, by (32), brings us to the stated result (25).  $\square$

As a corollary to Theorem 4.5, we obtain the following estimate for the error in the  $L^2(\Omega)$ -norm.

**Corollary 4.7.** *Consider the same premises as in Theorem 4.5 and assume that the dual regularity estimate (23) holds. Then, there exists a constant  $C$  such that*

$$(34) \quad \|u - u_{h,p}\|_{L^2(\Omega)} \leq C \left( \frac{h}{p^{1/2}} + \frac{1+\theta}{\sqrt{\alpha}} \right) \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right)^{1/2} + R,$$

where  $t_K = \min(r_K, s_K)$ ,  $\mu_K = \min(p_K + 1, r_K, s_K)$  and  $R = o(\|u - u_{h,p}\|_+)$ . Moreover, if the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v}): \mathbb{R}^d \rightarrow \mathbb{R}^d$  is twice continuously differentiable everywhere in  $\mathbb{R}^d$ , then there exists a constant  $C$  such that

$$(35) \quad R \leq C \max_{K \in \mathcal{T}_h} \left( \frac{p_K^{3/2}}{h_K^{d/2}} \right) \left( \sum_{K \in \mathcal{T}_h} \frac{h_K^{2\mu_K-2}}{p_K^{2t_K-3}} \|u\|_{H^{t_K}(K)}^2 \right).$$

*Proof.* The result follows immediately from Theorem 4.5 by selecting  $\psi = u - u_{h,p}$  and subsequently applying the regularity estimate (23).  $\square$

Let us briefly discuss the error estimates presented in Theorem 4.5 and Corollary 4.7. For  $h/p$  sufficiently small, we observe that

$$J_\psi(u) - J_\psi(u_{h,p}) \leq C \left( \frac{h^{\mu+\lambda-2}}{p^{t+m-3}} + \frac{1+\theta}{\sqrt{\alpha}} \frac{h^{\mu-1}}{p^{t-3/2}} \right) \|u\|_{H^t(\Omega)} \|z\|_{H^m(\Omega)}$$

and

$$\|u - u_{h,p}\|_{L^2(\Omega)} \leq C \left( \frac{h^\mu}{p^{t-1}} + \frac{1+\theta}{\sqrt{\alpha}} \frac{h^{\mu-1}}{p^{t-3/2}} \right) \|u\|_{H^t(\Omega)},$$

where  $t = \min_{K \in \mathcal{T}_h}(t_K)$ ,  $m = \min_{K \in \mathcal{T}_h}(m_K)$ ,  $\mu = \min_{K \in \mathcal{T}_h}(\mu_K)$  and  $\lambda = \min_{K \in \mathcal{T}_h}(\lambda_K)$ . Accordingly, when  $\theta = -1$ , we find that both estimates are optimal in  $h$  and slightly suboptimal in  $p$ , by one order in  $p$ . On the other hand, when  $\theta \neq -1$ , we find that the estimates are suboptimal in both  $h$  and  $p$ , by a factor of respectively  $h^{\lambda-1}/p^{m-3/2}$  and  $h/p^{1/2}$ . This suboptimality can be attributed to a lack of dual consistency; see Lemma 4.6. We note that, for  $h/p$  sufficiently small, the above estimates are identical to those obtained for interior penalty DG approximations of linear elliptic problems; cf. [14, Theorem 4.4].

**Remark 4.8.** *For the proof of Theorem 4.5 and Corollary 4.7 we assumed that the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v}): \mathbb{R}^d \rightarrow \mathbb{R}^d$  is Fréchet differentiable everywhere in  $\mathbb{R}^d$ . This was done in order to ensure that the dual problem (22) is well defined. It is envisaged that, with some additional effort, this assumption can be avoided, for instance, by reformulating the dual problem based on a regularization of the map  $\mathbf{v} \mapsto \mathbf{A}(\mathbf{v}): \mathbb{R}^d \rightarrow \mathbb{R}^d$ , for example, by using the techniques in [17].*

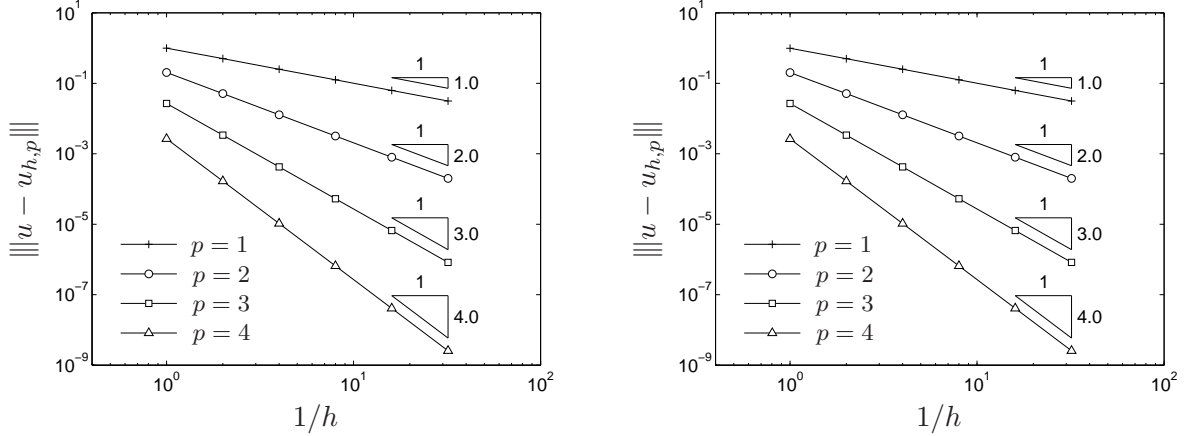
## 5 Numerical experiments

We present some numerical examples to verify the theoretical error estimates presented in Section 4. For simplicity, we restrict the presentation to 2D problems and consider uniformly refined meshes composed of affine quadrilaterals with uniform values of the polynomial degree  $\{p_K\}_{K \in \mathcal{T}_h}$ . Throughout this section, the interior penalty parameter is fixed at  $\alpha = 10$ . The nonlinear equations arising in the DG approximation are solved using an exact Newton method with a tolerance of  $10^{-10}$ . High-order numerical quadrature is used to integrate the terms appearing in the assembly of the associated algebraic system of equations, as well as to evaluate the error of the DG solution in various norms.

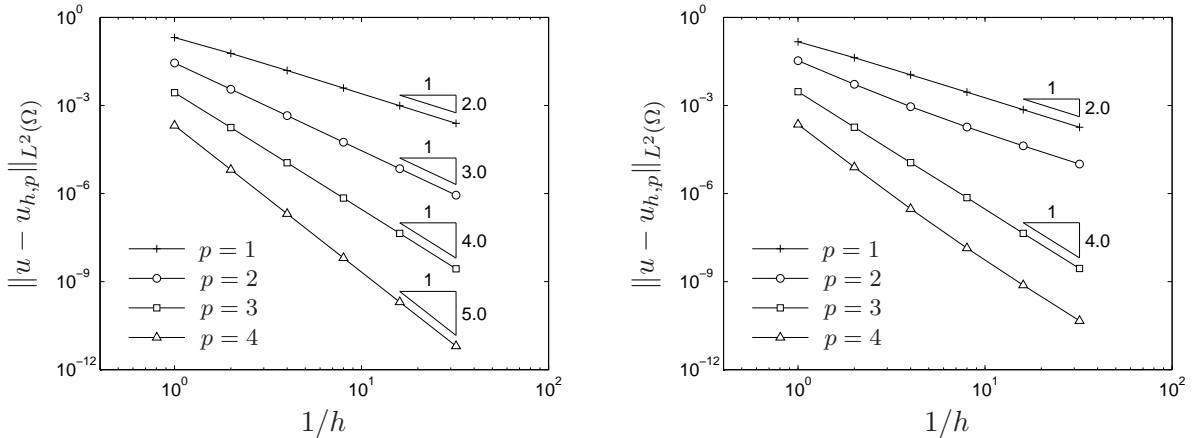
## 5.1 Example 1

For the first numerical example, we consider the problem of Example 1 in [4]; cf. also Example 1 in [15]. Accordingly, let  $\Omega = (-1, 1)^2$  with  $\Gamma_D = [-1, 1] \times \{-1\} \cup \{1\} \times [-1, 1]$  and  $\Gamma_N = [-1, 1] \times \{1\} \cup \{-1\} \times [-1, 1]$ , and let  $\mathbf{A}(\mathbf{x}, \nabla u) = (2 + (1 + |\nabla u|)^{-1}) \mathbf{I}$ , where  $\mathbf{I}$  is the  $2 \times 2$  identity matrix. The data  $f$ ,  $g_D$  and  $g_N$  are chosen such that the solution is given by the smooth function  $u(\mathbf{x}) = \cos(\pi x_1/2) \cos(\pi x_2/2)$ . We note that  $\mathbf{A}$  satisfies Assumption 1.1 with  $C_{\mathbf{A}} = 3$  and  $M_{\mathbf{A}} = 2$ .

We investigate the convergence of the DG approximation (13) on a sequence of successively refined meshes for different polynomial degrees. We consider two choices of the parameter  $\theta$ , viz.  $\theta = -1$  and  $\theta = 1$ . Figure 1 presents the convergence of the DG-norm of the error with  $h$ -refinement for  $p = 1, 2, 3$  and  $4$ . We observe that  $\|u - u_{h,p}\|$  converges to zero, for each fixed value of  $p$ , at a rate  $\mathcal{O}(h^p)$  as  $h \rightarrow 0$ . We note that these results are in perfect agreement with the theoretical error estimate presented in Theorem 4.3, and that the computed errors are virtually indistinguishable between the two choices of the parameter  $\theta$ . In Figure 2, we show the convergence of the  $L^2(\Omega)$ -norm of the error with  $h$ -refinement for  $p = 1, 2, 3$  and  $4$ . Here, significant differences are observed between the two choices of  $\theta$ . For  $\theta = -1$ , optimal convergence rates are obtained for all values of  $p$ ; i.e.,  $\|u - u_{h,p}\|_{L^2(\Omega)} = \mathcal{O}(h^{p+1})$  as  $h \rightarrow 0$  for each fixed value of  $p$ . For  $\theta = 1$  on the other



**Figure 1:** Example 1. Convergence of  $\|u - u_{h,p}\|$  with  $h$ -refinement for  $p = 1, 2, 3$  and  $4$ . Left:  $\theta = -1$ . Right:  $\theta = 1$ .



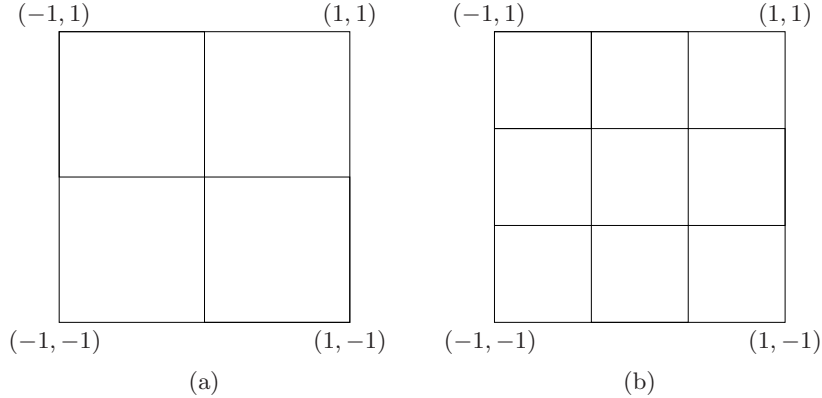
**Figure 2:** Example 1. Convergence of  $\|u - u_{h,p}\|_{L^2(\Omega)}$  with  $h$ -refinement for  $p = 1, 2, 3$  and  $4$ . Left:  $\theta = -1$ . Right:  $\theta = 1$ .

hand, we see that  $\|u - u_{h,p}\|_{L^2(\Omega)}$  behaves like  $\mathcal{O}(h^{p+1})$  as  $h \rightarrow 0$  for odd values of  $p$ , and like  $\mathcal{O}(h^p)$  as  $h \rightarrow 0$  for even values of  $p$ . This suboptimal convergence behavior for  $\theta = 1$  is attributable to a lack of dual consistency; cf. Lemma 4.6. The obtained convergence rates for  $\|u - u_{h,p}\|_{L^2(\Omega)}$  are in agreement with the theoretical error estimates presented in Corollary 4.7. Comparing the current result to the results reported for the same example in [15], we note that presented DG method with  $\theta = -1$  shows improved convergence behavior with respect to the error in the  $L^2(\Omega)$ -norm.

## 5.2 Example 2

In the second example, we consider a problem with a non-smooth solution. Let  $\Omega = (-1, 1)^2$  with  $\Gamma_D = \partial\Omega$ , and  $\mathbf{A}(\mathbf{x}, \nabla u) = (1 + e^{-|\nabla u|^2})\mathbf{I}$ , where  $\mathbf{I}$  denotes again the  $2 \times 2$  identity matrix. It is easy to verify that Assumption 1.1 is satisfied with  $C_{\mathbf{A}} = 1$  and  $M_{\mathbf{A}} = 1 - \sqrt{2/e}$ . The data  $f$  and  $g_D$  are chosen such that the solution is given by  $u(\mathbf{x}) = |\mathbf{x}|^3$ . We note that the solution features a singularity at the point  $(0, 0)$ , and that  $u \in H^{4-\epsilon}(\Omega)$  for arbitrary small  $\epsilon > 0$ .

We investigate the convergence behavior with  $p$ -refinement for the two meshes displayed in Figure 3. In Tables 1 and 2, we show the convergence of the DG-norm of the error and the  $L^2(\Omega)$ -norm for  $p = 1, 2, \dots, 24$ , and  $\theta = -1$ , grouped in odd and even values of  $p$ . For mesh (a), we observe that  $\|u - u_{h,p}\|$  converges at a rate of almost  $\mathcal{O}(p^{-6})$  as  $p \rightarrow \infty$ , and that  $\|u - u_{h,p}\|_{L^2(\Omega)}$  converges at a rate of approximately  $\mathcal{O}(p^{-15/2})$ . Comparing with the theoretical error estimates of Theorem 4.3 and Corollary 4.7, we note that these convergence rates are more than twice the predicted rate. Indeed, since  $u \in H^{4-\epsilon}$  for any  $\epsilon > 0$ , the expected convergence rates are  $\mathcal{O}(p^{-5/2+\epsilon})$  for  $\|u - u_{h,p}\|$  and  $\mathcal{O}(p^{-3+\epsilon})$  for  $\|u - u_{h,p}\|_{L^2(\Omega)}$ . This order-doubling convergence behavior is attributable to the fact that the singularity in  $u$  at the point  $(0, 0)$  coincides with a vertex of mesh (a). In the presence of such corner singularities, it is possible to establish *a priori* error estimates that reflect this order-doubling phenomenon by using approximation results in terms of weighted Sobolev norms; cf., for example, [16, Remark 3.8]. For mesh (b), on the other hand, the singularity in  $u$  lies in the interior of an element rather than at a vertex. Here, we see that the  $p$ -convergence rates approach the theoretical convergence rates predicted by Theorem 4.3 and Corollary 4.7. Indeed, it is found that  $\|u - u_{h,p}\|$  and  $\|u - u_{h,p}\|_{L^2(\Omega)}$  both behave like  $\mathcal{O}(p^{-3})$  as  $p \rightarrow \infty$ . For  $\|u - u_{h,p}\|$ , this constitutes a slight improvement of the theoretical convergence rate, by half an order in  $p$ , while for  $\|u - u_{h,p}\|_{L^2(\Omega)}$  the convergence rate is in perfect agreement. We end this example by stating that the results for  $\theta = 1$  are almost identical.



**Figure 3:** The two meshes considered for Example 2.

$p$	$\ u - u_{h,p}\ $		$\ u - u_{h,p}\ _{L^2(\Omega)}$	
1	3.11E+00	—	4.46E-01	—
3	4.09E-02	(3.94)	3.30E-03	(4.47)
5	2.17E-03	(5.75)	1.51E-04	(6.03)
7	2.84E-04	(6.05)	1.35E-05	(7.18)
9	6.38E-05	(5.94)	1.97E-06	(7.66)
11	1.96E-05	(5.89)	4.51E-07	(7.34)
13	7.32E-06	(5.88)	1.31E-07	(7.43)
15	3.16E-06	(5.88)	4.50E-08	(7.44)
17	1.51E-06	(5.88)	1.76E-08	(7.50)
19	7.86E-07	(5.88)	7.64E-09	(7.51)
21	4.36E-07	(5.88)	3.59E-09	(7.55)
23	2.55E-07	(5.89)	1.80E-09	(7.57)

$p$	$\ u - u_{h,p}\ $		$\ u - u_{h,p}\ _{L^2(\Omega)}$	
2	5.74E-01	—	8.60E-02	—
4	8.72E-03	(6.04)	6.74E-04	(7.00)
6	7.12E-04	(6.18)	3.71E-05	(7.15)
8	1.28E-04	(5.96)	5.00E-06	(6.97)
10	3.43E-05	(5.91)	9.28E-07	(7.55)
12	1.17E-05	(5.89)	2.37E-07	(7.49)
14	4.73E-06	(5.88)	7.52E-08	(7.45)
16	2.16E-06	(5.88)	2.78E-08	(7.46)
18	1.08E-06	(5.88)	1.15E-08	(7.49)
20	5.81E-07	(5.88)	5.19E-09	(7.54)
22	3.32E-07	(5.88)	2.53E-09	(7.56)
24	1.99E-07	(5.89)	1.30E-09	(7.59)

**Table 1:** Example 2. Convergence of  $\|u - u_{h,p}\|$  and  $\|u - u_{h,p}\|_{L^2(\Omega)}$  with  $p$ -refinement for mesh (a) and  $\theta = -1$ . The results are grouped in odd and even values of  $p$ . The quantities in brackets indicate the  $p$ -convergence rates.

$p$	$\ u - u_{h,p}\ $		$\ u - u_{h,p}\ _{L^2(\Omega)}$	
1	2.07E+00	—	2.31E-01	—
3	3.39E-02	(3.74)	2.22E-03	(4.23)
5	3.42E-03	(4.49)	1.67E-04	(5.07)
7	1.03E-03	(3.55)	4.39E-05	(3.96)
9	4.49E-04	(3.32)	1.81E-05	(3.53)
11	2.35E-04	(3.23)	9.29E-06	(3.32)
13	1.38E-04	(3.17)	5.44E-06	(3.20)
15	8.82E-05	(3.14)	3.48E-06	(3.13)
17	5.97E-05	(3.11)	2.36E-06	(3.09)
19	4.23E-05	(3.10)	1.68E-06	(3.06)
21	3.11E-05	(3.08)	1.24E-06	(3.04)
23	2.35E-05	(3.08)	9.42E-07	(3.02)

$p$	$\ u - u_{h,p}\ $		$\ u - u_{h,p}\ _{L^2(\Omega)}$	
2	2.21E-01	—	2.12E-02	—
4	3.63E-03	(5.93)	4.62E-04	(5.52)
6	1.09E-03	(2.96)	1.50E-04	(2.78)
8	4.75E-04	(2.90)	6.64E-05	(2.83)
10	2.49E-04	(2.89)	3.50E-05	(2.86)
12	1.47E-04	(2.90)	2.07E-05	(2.89)
14	9.38E-05	(2.91)	1.32E-05	(2.90)
16	6.36E-05	(2.92)	8.97E-06	(2.91)
18	4.51E-05	(2.92)	6.36E-06	(2.92)
20	3.31E-05	(2.93)	4.67E-06	(2.93)
22	2.50E-05	(2.93)	3.53E-06	(2.94)
24	1.94E-05	(2.94)	2.73E-06	(2.94)

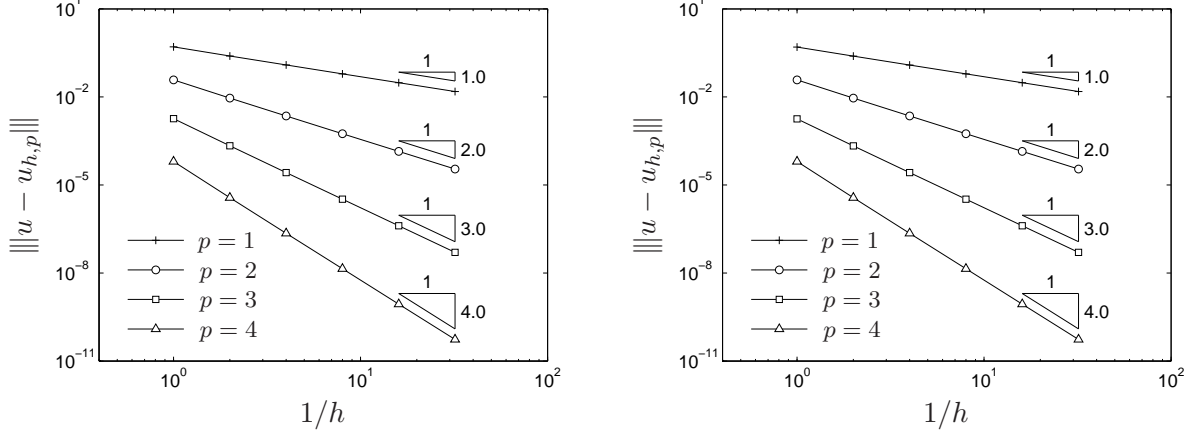
**Table 2:** Example 2. Convergence of  $\|u - u_{h,p}\|$  and  $\|u - u_{h,p}\|_{L^2(\Omega)}$  with  $p$ -refinement for mesh (b) and  $\theta = -1$ . The results are grouped in odd and even values of  $p$ . The quantities in brackets indicate the  $p$ -convergence rates.

### 5.3 Example 3

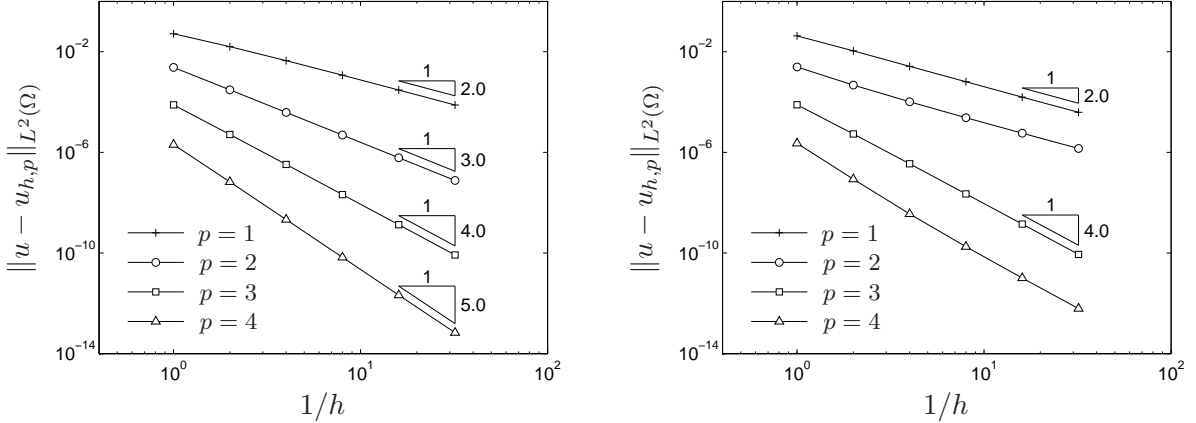
In the third and final example, we consider a case not fully covered by our theory. We consider the solution of the  $p(\mathbf{x})$ -Laplace equation with  $\mathbf{A}(\mathbf{x}, \nabla u) = |\nabla u|^{p(\mathbf{x})-2} \mathbf{I}$ , where  $p(\mathbf{x}) = 4 - |\mathbf{x}|^2$ . Note that  $\mathbf{A}$  does not comply with Assumption 1.1 for  $|\mathbf{x}| < 1$ . The problem is posed on the  $L$ -shaped domain  $\Omega = (-1, 1)^2 \setminus [0, 1] \times (-1, 0]$  with  $\Gamma_N = [-1, 1] \times \{1\} \cup \{-1\} \times [-1, 1]$  and  $\Gamma_D = \partial\Omega \setminus \Gamma_N$ . The data  $f$ ,  $g_D$  and  $g_N$  are chosen such that the solution is given by the smooth function  $u(\mathbf{x}) = x_1 e^{x_1 x_2}$ .

In Figure 4, we show the convergence of the DG-norm of the error with  $h$ -refinement for  $p = 1, 2, 3, 4$  and  $\theta = -1, 1$ . As in Example 1, we observe that  $\|u - u_{h,p}\|$  converges to zero, for each fixed value of  $p$ , at a rate  $\mathcal{O}(h^p)$  as  $h \rightarrow 0$ . Note that this is in perfect agreement with the theoretical error estimate presented in Theorem 4.3, even though the underlying Assumption 1.1 is not met. Also note that the results are virtually distinguishable between the two choices of the parameter  $\theta$ . In Figure 5, we present the convergence of the  $L^2(\Omega)$ -norm with  $h$ -refinement for  $p = 1, 2, 3, 4$  and  $\theta = -1, 1$ . Here, as in Example 1, significant differences are observed between the two values of  $\theta$ . For  $\theta = -1$ , optimal

convergence rates are obtained for all values of  $p$ ; i.e.,  $\|u - u_{h,p}\|_{L^2(\Omega)} = \mathcal{O}(h^{p+1})$  as  $h \rightarrow 0$  for each fixed value of  $p$ . For  $\theta = 1$  on the other hand, we see that  $\|u - u_{h,p}\|_{L^2(\Omega)}$  behaves like  $\mathcal{O}(h^{p+1})$  as  $h \rightarrow 0$  for odd values of  $p$ , and like  $\mathcal{O}(h^p)$  as  $h \rightarrow 0$  for even values of  $p$ . The convergence behavior for  $\|u - u_{h,p}\|_{L^2(\Omega)}$  is very similar to that seen in Example 1 and agrees well with the theoretical error estimates of Corollary 4.7.



**Figure 4:** Example 3. Convergence of  $|||u - u_{h,p}|||$  with  $h$ -refinement for  $p = 1, 2, 3$  and  $4$ . Left:  $\theta = -1$ . Right:  $\theta = 1$ .



**Figure 5:** Example 3. Convergence of  $\|u - u_{h,p}\|_{L^2(\Omega)}$  with  $h$ -refinement for  $p = 1, 2, 3$  and  $4$ . Left:  $\theta = -1$ . Right:  $\theta = 1$ .

## Acknowledgement

The work presented in this paper was completed while the author was a Ph.D. student at the Delft University of Technology working under the supervision of Dr. S. J. Hulshoff, for whose guidance and support the author is most grateful.

## A Nonlinear inf-sup theory

We include some auxiliary results regarding the well-posedness of nonlinear variational problems. Let  $U$  be a real Banach space equipped with the norm  $\|\cdot\|_U$ , and let  $V$  be a real reflexive Banach space equipped with the norm  $\|\cdot\|_V$ . We denote by  $U'$  and  $V'$  the



respective dual spaces, equipped with the norms

$$\|f\|_{U'} = \sup_{u \in U \setminus \{0\}} \frac{\langle f, u \rangle_{U', U}}{\|u\|_U}, \quad \|g\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{\langle g, v \rangle_{V', V}}{\|v\|_V},$$

where  $\langle \cdot, \cdot \rangle_{U', U}$  and  $\langle \cdot, \cdot \rangle_{V', V}$  are the duality pairings between  $U'$  and  $U$ , and  $V'$  and  $V$ , respectively.

The first result that we present constitutes a nonlinear extension of the classical well-posedness result of Banach, Nečas and Babuška; cf., for example, [6, Theorem 1.1]. The statement of the theorem and parts of its proof are adopted from [23, Appendix A], where the theorem is presented in a Hilbert space setting. We note that the theorem generalizes some other results from the literature; see, for example, [24, Theorem 25.B].

**Theorem A.1** (inf-sup conditions). *Let  $a: U \times V \rightarrow \mathbb{R}$  be a semilinear form, such that*

$$(36) \quad a(w_1; v) - a(w_2; v) \leq C_a \|w_1 - w_2\|_U \|v\|_V \quad \forall w_1, w_2 \in U, \forall v \in V$$

*for some constant  $C_a > 0$ . Then, the variational problem*

$$(37) \quad u \in U : \quad a(u; v) = f(v) \quad \forall v \in V$$

*admits a unique solution  $u \in U$  for every  $f \in V'$  if and only if*

$$(38) \quad \exists M_a > 0 : \quad \inf_{\substack{w_1, w_2 \in U \\ w_1 \neq w_2}} \sup_{v \in V \setminus \{0\}} \frac{a(w_1; v) - a(w_2; v)}{\|w_1 - w_2\|_U \|v\|_V} \geq M_a,$$

$$(39) \quad \sup_{w \in U} a(w; v) > 0 \quad \forall v \in V \setminus \{0\}.$$

*Moreover, for any  $g \in V' \setminus \{f\}$  and corresponding  $\tilde{u} \in U$  such that  $a(\tilde{u}; v) = g(v)$  for all  $v \in V$ , we have the following a priori estimate:*

$$(40) \quad \|u - \tilde{u}\|_U \leq \frac{1}{M_a} \|f - g\|_{V'}.$$

*Proof.* The proof proceeds in a similar manner as for the linear setting; cf., for example, [22]. For any fixed  $w \in U$ , consider the linear functional  $\phi_w: V \rightarrow \mathbb{R}$  of the form  $v \mapsto \phi_w(v) := a(w; v)$  for all  $v \in V$ . By virtue of (36) with  $w_1 = w$  and  $w_2 = 0$ , we have that

$$\|\phi_w\|_{V'} = \sup_{v \in V \setminus \{0\}} \frac{|\phi_w(v)|}{\|v\|_V} = \sup_{v \in V \setminus \{0\}} \frac{|a(w; v)|}{\|v\|_V} \leq C_a \|w\|_U.$$

Hence,  $\phi_w \in V'$ . Now, let  $A: U \rightarrow V'$  such that  $w \mapsto A(w) := \phi_w$  for all  $w \in U$ . The variational problem (37) is then equivalent to finding  $u \in U$  such that  $A(u) = f$  in  $V'$ . The existence and uniqueness of a solution  $u \in U$  is ensured if the operator  $A: U \rightarrow V'$  is injective and surjective.

Injectivity of  $A$  is established by verifying that  $A(w_1) = A(w_2)$  implies  $w_1 = w_2$ . By virtue of (38), we have that, for all  $w_1, w_2 \in U$ ,

$$\begin{aligned} \|A(w_1) - A(w_2)\|_{V'} &= \sup_{v \in V \setminus \{0\}} \frac{\langle A(w_1) - A(w_2), v \rangle_{V', V}}{\|v\|_V} \\ &= \sup_{v \in V \setminus \{0\}} \frac{a(w_1; v) - a(w_2; v)}{\|v\|_V} \\ (41) \quad &\geq M_a \|w_1 - w_2\|_U. \end{aligned}$$

Consequently,  $\|A(w_1) - A(w_2)\|_{V'} = 0$  implies  $\|w_1 - w_2\|_U = 0$ . Hence,  $A$  is injective.

Surjectivity of  $A$  is established by verifying that the range of  $A$ , hereafter denoted by  $\text{Im}(A)$ , coincides with  $V'$ . This is equivalent to showing that  $\text{Im}(A)$  is closed in  $V'$ , and that its orthogonal complement in  $V'$  is empty. To this end, let  $\{w_n\}_{n=0}^\infty$  be some sequence in  $U$  such that  $\{A(w_n)\}_{n=0}^\infty$  is a Cauchy sequence in  $V'$ . Then, from (41), it follows that  $\{w_n\}_{n=0}^\infty$  is Cauchy in  $U$ . Let  $w$  be its limit. On account of (36), we have that  $A(w_n) \rightarrow A(w)$  as  $n \rightarrow \infty$ ; indeed, for  $n \rightarrow \infty$ ,

$$\begin{aligned} \|A(w) - A(w_n)\|_{V'} &= \sup_{v \in V \setminus \{0\}} \frac{\langle A(w) - A(w_n), v \rangle_{V', V}}{\|v\|_V} \\ &= \sup_{v \in V \setminus \{0\}} \frac{a(w; v) - a(w_n; v)}{\|v\|_V} \\ &\leq C_a \|w - w_n\|_U \rightarrow 0. \end{aligned}$$

This in turn implies that  $A(w) \in \text{Im}(A)$  and, thus, that  $\text{Im}(A)$  is closed. It remains to show that the orthogonal complement of  $A$  in  $V'$  is empty. Let us argue by contradiction by supposing that  $\text{Im}(A) \subsetneq V'$ . Then, by the Hahn-Banach theorem in the form of [25, Proposition 3], there exists a  $v_0 \in V''$  such that  $\langle A(w), v_0 \rangle_{V', V''} = 0$  for every  $w \in U$ . Since  $V$  is reflexive, we can identify  $V''$  with  $V$  so that  $v_0 \in V$ . Accordingly, we have

$$0 = \langle A(w), v_0 \rangle_{V', V} = a(w; v_0) \quad \forall w \in U,$$

which is in contradiction to (39). This implies that  $V' \setminus \text{Im}(A) = \emptyset$  and, therefore,  $\text{Im}(A) \equiv V'$ . Hence,  $A$  is surjective.

Based on the above, we conclude that (37) has a unique solution  $u \in U$  for every  $f \in V'$  whenever (38) and (39) hold. The *a priori* estimate (40) readily follows by noting that, from (38) with  $w_1 = u$  and  $w_2 = \tilde{u}$ ,

$$M_a \|u - \tilde{u}\|_U \leq \sup_{v \in V \setminus \{0\}} \frac{a(u; v) - a(\tilde{u}; v)}{\|v\|_V} = \sup_{v \in V \setminus \{0\}} \frac{\langle f - g, v \rangle_{V', V}}{\|v\|_V} = \|f - g\|_{V'}.$$

It remains to prove that (38) and (39) are also necessary conditions for ensuring well-posedness of (37). The necessity of (38) follows from uniqueness. Indeed, assume that there exists a pair  $u_1, u_2 \in U$ ,  $u_1 \neq u_2$ , such that

$$\sup_{v \in V \setminus \{0\}} \frac{a(u_1; v) - a(u_2; v)}{\|v\|_V} = 0.$$

This would imply that  $a(u_1; v) = a(u_2; v)$  for every  $v \in V$ , which is in contradiction to uniqueness. The necessity of (39) follows from existence. To see this, assume that there exists some  $v_0 \in V \setminus \{0\}$  such that  $a(w; v_0) = 0$  for every  $w \in U$ . By the Hahn-Banach theorem, there exists an  $\tilde{f} \in V'$  such that  $\tilde{f}(v_0) \neq 0$ , implying

$$0 = a(w; v_0) = \tilde{f}(v_0) \neq 0,$$

which is a contradiction to the solvability of (37). This concludes the proof.  $\square$

The second result that we present provides equivalent inf-sup conditions. It constitutes a nonlinear extension of [18, Proposition A.2]. The result is not essential for the material presented in this paper, but is included nevertheless because it could be of independent interest.

**Theorem A.2.** Let  $a: U \times V \rightarrow \mathbb{R}$  be a semilinear form, such that

$$(42) \quad a(w_1; v) - a(w_2; v) \leq C_a \|w_1 - w_2\|_U \|v\|_V \quad \forall w_1, w_2 \in U, \forall v \in V$$

for some constant  $C_a > 0$ . Moreover, let the map  $w \mapsto a(w; \cdot)$  be everywhere Fréchet differentiable in  $U$ , and denote by  $a'(q; w, \cdot)$  the corresponding Fréchet derivative at  $q \in U$  in the direction  $w \in U$ . The following statements are equivalent, with identical constant  $M_a > 0$ .

(i) It holds that:

$$(43) \quad \exists M_a > 0 : \inf_{\substack{w_1, w_2 \in U \\ w_1 \neq w_2}} \sup_{v \in V \setminus \{0\}} \frac{a(w_1; v) - a(w_2; v)}{\|w_1 - w_2\|_U \|v\|_V} \geq M_a,$$

$$(44) \quad \sup_{w \in U} a(w; v) > 0 \quad \forall v \in V \setminus \{0\}.$$

(ii) For all  $q \in U$ , it holds that:

$$(45) \quad \inf_{v \in V \setminus \{0\}} \sup_{w \in U \setminus \{0\}} \frac{a'(q; w, v)}{\|w\|_U \|v\|_V} \geq M_a,$$

$$(46) \quad \sup_{v \in V} a'(q; w, v) > 0 \quad \forall w \in U \setminus \{0\}.$$

*Proof.* We first prove that (i) implies (ii). For arbitrary fixed  $v \in V$ , let  $J_v \in V'$  such that  $\|J_v\|_{V'} = 1$  and  $J_v(v) = \|v\|_V$ . The existence of such a linear functional  $J_v \in V'$  follows by application of the Hahn-Banach theorem; see, for example, [25, p. 5–6]. Then, for any  $q \in U$ , let  $u_{q,v} \in U$  be the solution of

$$a(u_{q,v}; z) = a(q; z) + \|v\|_V J_v(z) \quad \forall z \in V.$$

By Theorem A.1 and the premises (42)–(44), we have that the solution  $u_{q,v} \in U$  exists and is unique. Using (43) and recalling that  $\|J_v\|_{V'} = 1$ , we deduce the following estimate:

$$\|u_{q,v} - q\|_U \leq \frac{1}{M_a} \sup_{z \in V \setminus \{0\}} \frac{a(u_{q,v}; z) - a(q; z)}{\|z\|_V} = \frac{1}{M_a} \sup_{z \in V \setminus \{0\}} \frac{\|v\|_V J_v(z)}{\|z\|_V} = \frac{1}{M_a} \|v\|_V.$$

Hence, we have that

$$\inf_{v \in V \setminus \{0\}} \frac{a(u_{q,v}; v) - a(q; v)}{\|u_{q,v} - q\|_U \|v\|_V} = \inf_{v \in V \setminus \{0\}} \frac{J_v(v)}{\|u_{q,v} - q\|_U} = \inf_{v \in V \setminus \{0\}} \frac{\|v\|_V}{\|u_{q,v} - q\|_U} \geq M_a.$$

Noting that

$$\begin{aligned} \inf_{v \in V \setminus \{0\}} \sup_{w \in U \setminus \{0\}} \frac{a'(q; w, v)}{\|w\|_U \|v\|_V} &\geq \inf_{v \in V \setminus \{0\}} \sup_{w \in U \setminus \{0\}} \inf_{t > 0} \frac{a(q + tw; v) - a(q; v)}{t \|w\|_U \|v\|_V} \\ &\geq \inf_{v \in V \setminus \{0\}} \frac{a(u_{q,v}; v) - a(q; v)}{\|u_{q,v} - q\|_U \|v\|_V}, \end{aligned}$$

we then obtain (45). To show (46), let

$$v_0 = \arg \sup_{v \in V \setminus \{0\}} \left( \inf_{\substack{w_1, w_2 \in U \\ w_1 \neq w_2}} \frac{a(w_1; v) - a(w_2; v)}{\|w\|_U} \right).$$

By (43), we have that, for any  $w \in U \setminus \{0\}$ ,

$$a'(q; w, v_0) \geq \inf_{t>0} \frac{a(q + tw; v_0) - a(q; v_0)}{t} \geq \inf_{t>0} \frac{M_a \|tw\|_U \|v_0\|_V}{t} = M_a \|w\|_U \|v_0\|_V,$$

yielding

$$\sup_{v \in V} a'(q; w, v) \geq a'(q; w, v_0) \geq M_a \|w\|_U \|v_0\|_V > 0 \quad \forall w \in U \setminus \{0\}.$$

Hence, we have proved that (i) implies (ii).

To prove the reverse implication, consider an arbitrary fixed  $w \in U$ , and let  $J_w \in U'$  such that  $\|J_w\|_{U'} = 1$  and  $J_w(w) = \|w\|_U$ ; cf. again [25, p. 5–6]. Then, given any  $q \in U$ , let  $v_q \in V$  be the solution of

$$(47) \quad a'(q; y, v_q) = \|w\|_U J_w(y) \quad \forall y \in U.$$

By (42), we have that

$$a'(q; w, v) \leq C_a \|w\|_U \|v\|_V \quad \forall q, w \in U, \forall v \in V.$$

In view of this and the premises (45)–(46), existence and uniqueness of the solution  $v_q \in V$  to the problem (47) are asserted by the classical well-posedness result of Banach, Nečas and Babuška; cf., for example, [6, Theorem 1.1]. The following *a priori* estimate is derived:

$$\|v_q\|_V \leq \frac{1}{M_a} \sup_{y \in U \setminus \{0\}} \frac{a'(q; y, v_q)}{\|y\|} = \frac{1}{M_a} \sup_{y \in U \setminus \{0\}} \frac{\|w\|_U J_w(y)}{\|y\|} = \frac{1}{M_a} \|w\|_U.$$

Accordingly, we have that

$$\inf_{w \in U \setminus \{0\}} \frac{a'(q; w, v_q)}{\|w\|_U \|v_q\|_V} = \inf_{w \in U \setminus \{0\}} \frac{\|w\|_U J_w(w)}{\|w\|_U \|v_q\|_V} = \inf_{w \in U \setminus \{0\}} \frac{\|w\|_U}{\|v_q\|_V} \geq M_a.$$

By virtue of the mean-value theorem, we then arrive at the inequality

$$\inf_{\substack{w_1, w_2 \in U \\ w_1 \neq w_2}} \sup_{v \in V \setminus \{0\}} \frac{a(w_1; v) - a(w_2; v)}{\|w_1 - w_2\|_U \|v\|_V} \geq \inf_{\substack{w_1, w_2 \in U \\ w_1 \neq w_2}} \sup_{v \in V \setminus \{0\}} \inf_{q \in U} \frac{a'(q; w_1 - w_2; v)}{\|w_1 - w_2\|_U \|v\|_V} \geq M_a,$$

which is (43). Finally, to show (44), let

$$w_0 = \arg \sup_{w \in U \setminus \{0\}} \left( \inf_{q \in U} \inf_{v \in V \setminus \{0\}} \frac{a'(q; w, v)}{\|w\|_U} \right).$$

By virtue of the mean-value theorem, we have that

$$a(w_0; v) \geq \inf_{q \in U} a'(q; w_0, v) \geq M_a \|w_0\|_U \|v\|_V > 0 \quad \forall v \in V \setminus \{0\}$$

This concludes the proof.  $\square$

## References

- [1] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.
- [2] I. Babuška and M. Suri. The  $h$ - $p$  version of the finite element method with quasiuniform meshes. *Math. Model. Numer. Anal.*, 21:199–238, 1987.
- [3] C. Bi and Y. Lin. Discontinuous Galerkin method for monotone nonlinear elliptic problems. *Int. J. Numer. Anal. Mod.*, 9:999–1024, 2012.
- [4] R. Bustinza and G. N. Gatica. A local discontinuous Galerkin method for nonlinear diffusion problems with mixed boundary conditions. *SIAM J. Sci. Comput.*, 26:152–177, 2004.
- [5] P. G. Ciarlet and P.-A. Raviart. Interpolation theory of curved elements, with applications to finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 1:217–249, 1972.
- [6] D. A. Di Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*, volume 69 of *Mathématiques et Applications*. Springer-Verlag, 2012.
- [7] V. Dolejší. Analysis and application of the IIPG method to quasilinear nonstationary convection-diffusion problems. *J. Comput. Appl. Math.*, 222:251–273, 2008.
- [8] V. Dolejší, M. Feistauer, and V. Sobotíková. Analysis of the discontinuous Galerkin method for nonlinear convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 194:2709–2733, 2005.
- [9] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, 1992.
- [10] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, second edition, 1983.
- [11] T. Gudi, N. Nataraj, and A. K. Pani.  $hp$ -discontinuous Galerkin methods for strongly nonlinear elliptic boundary value problems. *Numer. Math.*, 109:233–268, 2008.
- [12] T. Gudi, N. Nataraj, and A. K. Pani. An  $hp$ -local discontinuous Galerkin method for some quasilinear elliptic boundary value problems of nonmonotone type. *Math. Comp.*, 77:731–656, 2008.
- [13] T. Gudi and A. K. Pani. Discontinuous Galerkin methods for quasi-linear elliptic problems of nonmonotone type. *SIAM J. Numer. Anal.*, 45:163–192, 2007.
- [14] K. Harriman, P. Houston, B. Senior, and E. Süli.  $hp$ -version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form. In S.-Y. Cheng, C.-W. Shu, and T. Tang, editors, *Recent Advances in Scientific Computing and Partial Differential Equations*, volume 330 of *Contemporary Mathematics*, pages 89–119. AMS, 2003.
- [15] P. Houston, J. A. Robson, and E. Süli. Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems I: The scalar case. *IMA J. Numer. Anal.*, 25:726–749, 2005.

- [16] P. Houston, C. Schwab, and E. Süli. Discontinuous  $hp$ -finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39:2133–2163, 2002.
- [17] J. M. Lasry and P. L. Lions. A remark on regularization in Hilbert spaces. *Israel Math. J.*, 55:257–266, 1986.
- [18] J. M. Melenk and C. Schwab. An  $hp$  finite element method for convection-diffusion problems. Technical Report 97-05, ETH Zürich, 1997.
- [19] C. Ortner and E. Süli. Discontinuous Galerkin finite element approximation of nonlinear second-order elliptic and hyperbolic systems. *SIAM J. Numer. Anal.*, 45:1370–1397, 2007.
- [20] A. Quarteroni. Some results of Bernstein and Jackson type for polynomial approximation in  $l^p$ -spaces. *Japan J. Appl. Math.*, 1:173–181, 1984.
- [21] B. Rivière, M. F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.*, 3:337–360, 1999.
- [22] C. Schwab. *p- and hp-Finite Element Methods: Theory and Applications in Solid and Fluid Mechanics*. Oxford University Press, 1998.
- [23] E. H. van Brummelen and R. de Borst. On the nonnormality of subiteration for a fluid-structure-interaction problem. *SIAM J. Sci. Comput.*, 27:599–621, 2005.
- [24] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/B: Nonlinear Monotone Operators*. Springer-Verlag, 1990.
- [25] E. Zeidler. *Applied Functional Analysis: Main Principles and Their Applications*, volume 109 of *Applied Mathematical Sciences*. Springer-Verlag, 1995.